Functional Signatures in Protein-protein Interactions and Their Impact on Signaling Pathways

A Thesis

Submitted to the Faculty

of

Drexel University

by

Yichuan Liu

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

June 2010

TABLE OF CONTENTS

Contents.....i

Acknowledgements……iv

List of Tables…….....v

List of Figures…….....vi

Abstract…….....viii

Chapter 1: Objectives and Overview.....1

1.1 Transition stage for protein-protein interactions……1

1.2 Functional signatures and databases……2

1.3 Case study selections……5

1.4 Current methods for PPI predictions…….....9

1.4.1 Primary structure and associated information……....9

1.4.2 Gene Ontology (GO)……..10

1.4.3 Geometric simulations for binding interfaces…….12

1.4.4 Functional signatures……..13

1.5 SNP alter signatures……15

1.6 Current methods in SNP-phenotype association…….17

1.6.1 Critical amino acid alterations…….……18

1.6.2 Data-mining/Computational Intelligence methods……19

Chapter 2: Modular composition (Domain/Motif) signature strings predicts kinase/substrate interactions.....23

2.1 Summary.....23

2.2 Introduction.....24

2.3 Method.....27

2.3.1 PPI data for phosphorylation events…….27

2.3.2 Scanning proteins for PROSITE domains and their enrichment in protein subgroups…….....28

2.3.3 Statistical enrichment of domains in protein subgroups…...28

2.3.4 Score matrix for signature pairs in PPIs…….29

2.3.5 Prediction accuracy for string pairs……..30

2.3.6 Sensitivity, specificity, precision, and recall……...30

2.3.7 Validation with independent datasets…….…….....31

2.3.8 Comparison with other computational models…..……....32

2.4 Results ……...……33

2.4.1 PROSITE domains enriched in kinase and their substrates……33

2.4.2 Score matrices for identifying domain signature sets enriched in known kinase protein interactions…….35

2.4.3 Validation with independent experimental datasets…….39

2.4.4 Matching kinase with substrates in expanding Previously annotated cellular pathways…….....41

2.5 Discussion & Conclusion.....43

Chapter 3: YiRen: A prediction tool for protein binding interactions based on functional domain pair enrichment.....47

3.1 Summary.....47

3.2 Background…….....48

3.3 Implementations.....51

3.4 Results & Discussion.....53

3.5 Availability.....55

Chapter 4: Domain altering SNPs in the human proteome and their impact on signaling pathways…….....56

4.1 Summary.....…...56

4.2 Background….....57

4.3. Methods…….….....59

4.3.1 Discovery of domain-altering SNPs.....59

4.3.2 Bonds broken between a protein with a domain altering SNP and its neighbors in signaling pathways ……60

4.4 Results…….……..61

4.5 Discussion & Conclusions..……..73

Chapter 5: Future Work……....76

5.1 Overview……76

5.2 Tag SNPs among different populations ……...79

5.3 Functional signatures in cross-specifies talks ……...83

Appendices……88

List of References.....101

Vita……...102

ACKNOWLEDGEMENTS

I would like to thank my Ph.D. advisor, Dr. Aydin Tozeren, for his guidance, encouragement and support during the development of my entire Ph.D work. He is an energetic advisor and one of the smartest people I have ever met. His enthusiasm for scientific research was the reason why I decided to pursue a career in the research field. I want to thank him for teaching me how to face the challenges, solve the problems and persist to accomplish the goals of the past three years. What I learned from him in the past three years is a wealth of knowledge for my future career path. Furthermore, I really appreciate his generous help during my life living in United States.

I would like to thank the members of my Ph.D. committee for their very helpful suggestions and comments through the development of my thesis. These members include Dr. Andres Kriete, Dr. Hualou Liang, Dr. Andrew Quong and Dr. Lyle Ungar.

I also much appreciate the helps offered from my current and previous lab mates, William Dampier, Perry Evans, Noor Dawany, Mahdi Sarmady, Adam Ertel and Mike Gormley. Many thanks to them for their valuable suggestions during my Ph.D study. I gained much knowledge from my discussions with them and it was my pleasure to have had partners like them.

LIST OF TABLES

Table 2.1 Accuracy and coverage of the present approach for predicting kinase – substrate interactions.……38

Table 2.2 Efficiency of the present score matrix enrichment in matching known phosphorylation PPIs..……..41

Table 3.1: validation of YiRen prediction across independent PPI set with other methods for 37 PPIs in new 2009 version from 54 proteins pool……..……54

Table 4.1: Statistical enrichment of domain altering SNPs in the OMIM database…….…….64

Table 4.2 List of proteins with domain altering SNPs ……..65

Table 4.3: The top ten most highly connected proteins with domain altering SNPs ……..73

Table 5.1 Examples of population-specific tag SNPs and corresponding phenotypes…….82

Table 5.2 Angles of ELM ligand binding motif vectors across species…….....85

LIST OF FIGURES

Figure 1.1 Single domain distribution of human proteins……...3

Figure 1.2 Random examples of four protein-protein interactions (PPIs)……..4

Figure 1.3 Network of PROSITE pairs related to Protein Kinase Activity……...5

Figure 1.4 Model for the kinase/substrate interactions …….…...6

Figure 1.5 Kinase related diseases in AGC group…….7

Figure 1.6 Examples of Domain-altering SNPs……17

Figure 2.1 Protein domains enriched in kinase- and substrate subtypes……34

Figure 2.2 Domain number density for kinases and their substrates…….35

Figure 2.3 Heat maps of the domain signature pairs associated with kinase-substrate interactions……..37

Figure 2.4 KEGG MAPK pathway revised by adding predicted phosphorylation events…….42

Figure 3.1 Flow chart of YiRen PPI prediction tool……..50

Figure 4.1 Alteration of 3D structure of TP53 due to presence of SNP rs28934571…….63

Figure 4.2 Statistically enriched GO and KEGG……...65

Figure 4.3 Proteins with domain altering SNPs on KEGG pathways ……...68

Figure 4.4 Statistical enrichment of signature pairs in PPIs involved with DA-SNPs & portion of broken edges……...71

Figure 5.1 Connectivity maps of the DDI interactions……77

Figure 5.2 Distribution of SNPs among different populations……..80

Figure 5.3 Distribution of tagSNPs that are related to diseases/disorders…….81

Figure 5.4 Frequencies of 81 ligand binding motifs of humans and hepatitis B……84

Figure 5.5 Frequencies of PROSITE domain and ELM motifs across species.....85

ABSTRACT

Functional Signatures in Protein-protein Interaction and Their Impact on Signaling Pathways

Yichuan Liu

Aydin Tozeren Ph.D.

Protein-protein interactions (PPIs) are the most fundamental biological processes at the molecular level. PPIs have been proved to be involved in pathologic mechanisms of many diseases. The experimental methods for testing the binding of PPIs are time-consuming and limited by analogs for many reactions. As a result, a computational model is necessary to predict PPIs and to explore the consequences of signal alterations in biological pathways.

A score matrix selection model was built based on overrepresented signature combinations. The case study focused on phosphorylation, which is a well studied post-translational modification category. The signature pairs were extended to signature-string pairs because of the multiple binding sites of kinase/substring interactions. A hypergeometric test was applied to select the significant signals due to the multiple-multiple relationship between the proteins and the domains/motifs. The prediction result shows an extremely high specificity (~100% compared to random combinations in the human protein pool) and an acceptable sensitivity rate (>65%) according to 10-fold evaluations. The score matrix model has then been extended to the user-defined-input software, named

'YiRen'. A group of PPIs related to transcription factors were evaluated in the test case.

Since the signatures embedded in protein sequences effect signal strength and they could be applied as the predictors in PPIs, alterations of these signatures could lead to broken edges in biological networks. An SNP is a kind of sequence variation. It is the major cause of human genetic variations and plays a key role in personalized medicine. In the DA-SNP (Domain-altering SNP) model, the SNPs from a dbSNP database were filtered through the domain regions on human proteomes. The SNPs were selected if they altered the domain signal strength by more than 10%. Then the selected SNPs were checked through an OMIM database for SNP-disease mappings, while the SNP-corresponding proteins were checked through the protein-disease database in Human Protein Reference Database (HPRD). The altered domains then projected into significant signature vectors in PPI prediction and the broken edges in biological pathways. The model linked the phenotypes and the sequence variation together with functional units in order to provide potential explanations for the phenotypes.

CHAPTER 1: OBJECTIVES AND OVERVIEW

1.1 Transition stage for protein-protein interactions

The transition stage for protein binding is a period for breaking old bonds and forming new ones [1-2]. The formation could physically change the shapes of the protein and alter the activation energy for the chemical processes [3]. The binding energy can be determined and provides the enzyme specificity for the catalysis process [2, 4]. The protein-protein interaction (PPI) processes in the transition stage are extremely important because in short period binding, they will dramatically change the formation of the target proteins, and then cause variations throughout the whole system [2]. Typical transition binding protein-protein interactions include phosphorylation and transcription factor activity. Phosphorylation is a typical catalysis process, which adds a phosphate (PO4) group to a protein [5]. The addition of the phosphate to Ser, Tyr, and Thr residues could cause significant differences in the protein 3D structure, if the modified site responds to critical regions [2]. Kinase phosphorylation processes have proven that they play a critical role in cancer research and drug development [6-7]. Transcription factors always combined with other proteins to enhance or inhibit the transcription biological processes [8]. In eukaryotes, many transcription factors do not bind to DNA directly, but directly interact with RNA polymerase [9], which leads to protein-protein interactions. Transcription factor activity is

important because it is involved in the development of organisms [10], such as sex-determined region Y regulations [11]. Transcription factor activity also responds to the environment via signals, such as changes in temperature [12] and oxygen levels [13].

1.2 Functional signatures and databases

The functional signatures in the protein sequences can be categorized by protein domain structures and linear motifs. The protein domain is a part of the protein sequence and structure, which can evolve, function, and exist independently of the rest of the protein chain [14]. The length of the domains is variant from about 25 amino acids to 500 amino acids, which are on the average of 100 amino acids long [15]. The longest domain unit is lipoxygenase-1, with 692 residues [16]. Protein domains often form functional units such as the calcium-binding EF hand domain of calmodulin [17]. Protein domains are often found in the protein primary sequence, secondary structure or tertiary structure. The database of protein domains, families and functional sites (PROSITE) [18] is a database containing the domain information in grammar forms for the domains in primary and secondary structures. Grammar forms indicate that the domain is represented by a regular expression (pattern) or a position weight matrix (profile). The search engine of PROSITE [19] can be downloaded and takes the primary sequence of the protein as inputs and returns the hits of the domains found in the definition database. The distributions of the PROSITE domains across the human protein sequences are not uniform with low entropy (Figure 1.1).

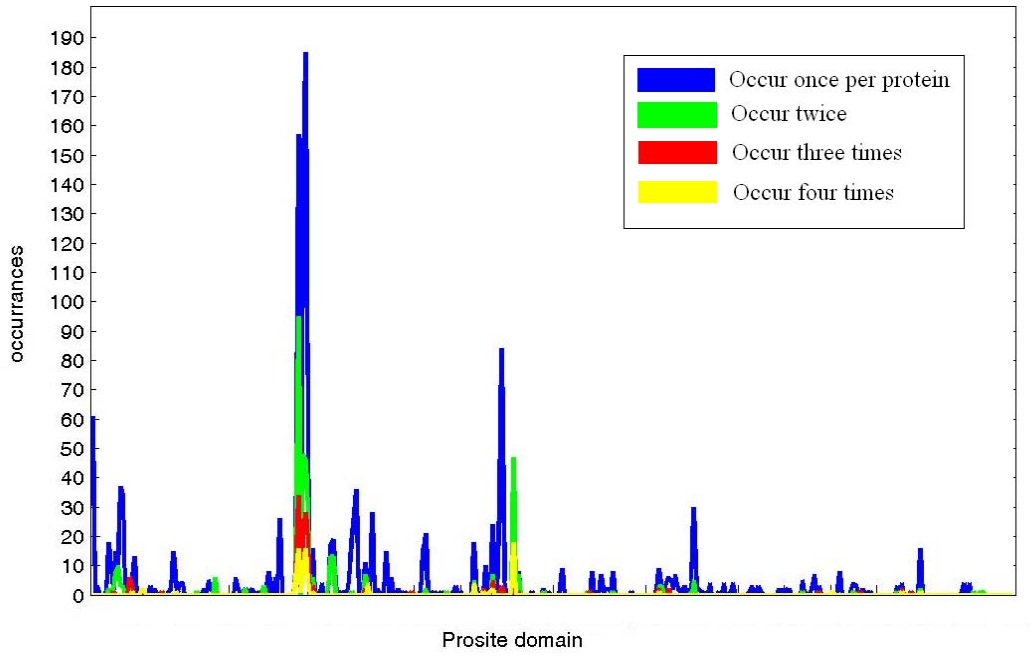


Figure 1.1 Single domain distribution of human proteins

Occurrences of proteins which contain certain frequency of PROSITE domain.

The tertiary structure (3D) of the protein domains is useful in defining protein/domain families. Molecular evolution gives rise to families of related proteins with a similar sequence and structure. However, sequence similarities can be extremely low between proteins that share the same structure [20]. The current 3D database related to the protein 3D domains is the Protein Data Bank (PDB) [21], which contains over 45,000 experimentally determined protein structures. As shown in Figure 1.2, several examples illustrate the protein-protein binding complex involved in domain-domain or domain-motif interactions. The problem with the 3D structure is coverage and incomplete structures (fragments). The binding visualizations were done by ZDOCK [22].

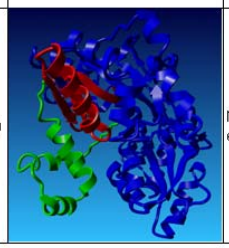
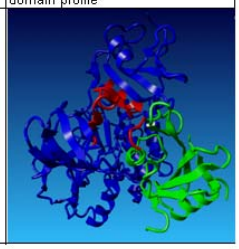
DDI involved	Laminin-type EGF-like (LE) domain signature	Homeobox' domain profile	Ankyrin repeat profile	Src homology 3 (SH3) domain profile; Src homology 2 (SH2) domain profile
Docking views	No 3D structure (only fragments)		No docking predictions (PDB file exceed the limited size)	

Figure 1.2 Random examples of four protein-protein interactions (PPIs)

Four examples which are involved in domain-domain interactions (DDIs) were selected from human protein reference database (HPRD). The protein A and B in the interaction were colored as blue and green, the domain region for the bindings were colored as red. Two of the examples do not have the visualizations due to the fragments of the 3D structures.

The motifs are sequence motifs and/or structural motifs. Sequence motifs are nucleotide or amino-acid sequence patterns that are widespread and have a biological significance [14]. The sequence motifs are distinguished from structural motifs by way of the proteins. The structural motifs are formed by the three dimensional arrangement of amino acids, which may not be adjacent [14]. Compared to the domains that have multiple levels of structures, protein motifs are more linear, and the size of the motifs are relatively small compared to the domains (3~12 amino acids) [23]. The motif dictionaries that were applied to the present study are the PROSITE [18] and Eukaryotic Linear Motif (ELM) [24]. ELM treats motifs as regular expressions. Other famous functional unit databases related to motifs include Pfam [25] and Interpro [26].

Another reason for choosing the domain/motif functional signature is that the interactions of any proteins can be converted to a network composed of functional signatures. In other words, functional signatures are generalized versions of amino acid sequences. An example is shown in Figure 1.3. The PPIs network relates to the protein kinase activities (PKA) that are defined by the Gene Ontology (GO)

database [27]. The network contains 1048 protein interactions with 497 proteins. The PPIs network is converted to the functional network, which then contains 187 interactions that are related to 41 functional signatures.

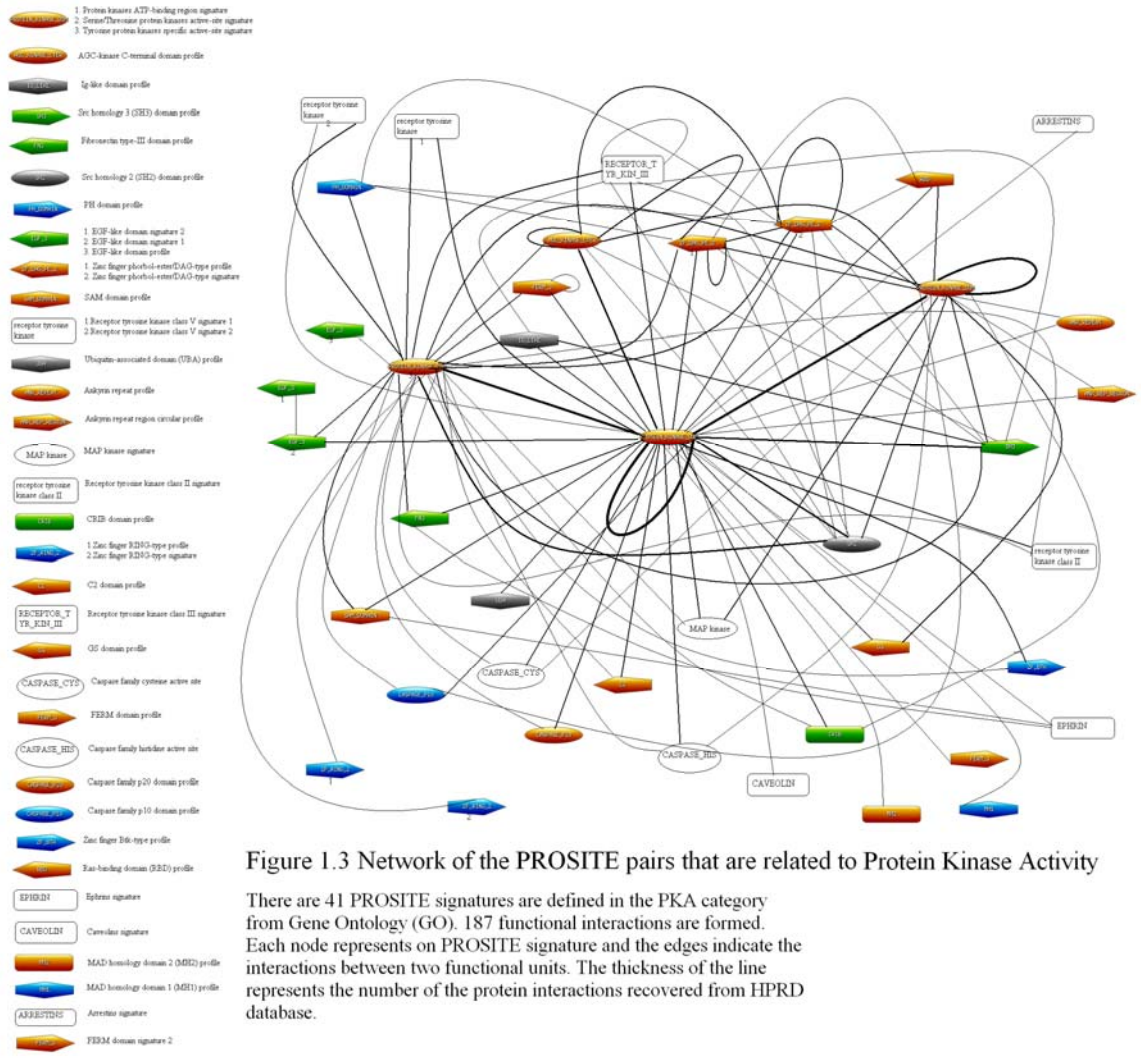


Figure 1.3 Network of the PROSITE pairs that are related to Protein Kinase Activity

There are 41 PROSITE signatures are defined in the PKA category from Gene Ontology (GO). 187 functional interactions are formed. Each node represents on PROSITE signature and the edges indicate the interactions between two functional units. The thickness of the line represents the number of the protein interactions recovered from HPRD database.

1.3 Case study selections

The protein interaction groups must satisfy the following conditions to be considered as ideal candidates. 1) The number of PPIs should be in the hundreds, since most statistical analysis methods prefer a large sample size. 2) The PPIs have proven that their bindings are involved with the domain/motif functional

signatures. Actually, there are many types of PPIs, and the domain-domain interactions (DDI) or domain-motif interactions (DMI) are only one group among different PPI types. 3) The PPIs relate to complex diseases, which provide significance to the project. In addition to the above requirements, it is better for the candidate groups to have experimental interaction databases to avoid biases and approximation errors during the selection process.

Two case studies related to the transition binding were selected for the thesis: phosphorylation events and transcription factor activities. Phosphorylation events are a well-studied group of PPIs that relate to the domain/motif interactions (Figure 1.4) [14].

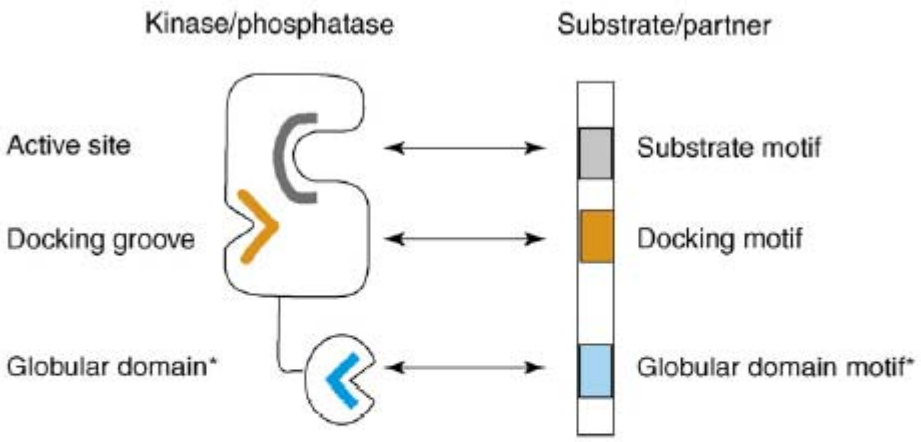


Figure 1.4 Model for the kinase/substrate interactions

An example of kinase/substrate binding that is involved with multiple domains/motifs.

The databases such as Post-Translational Modification (PTM) [28] and PhosphoELM [29] were the selections of experimentally determined phosphorylation events based on PubMed. These databases can be accessed and

downloaded directly from the Internet. Both databases contain enough PPI numbers for analysis. The PTM (2006 version) contains over 5,000 phosphorylation events [28] and the PhosphoELM contain about 3,000 phosphorylation events [29]. As mentioned in section 1.1, the phosphorylation events are considered as the causes/co-effectors in many genetic diseases/disorders. The kinases and their corresponding substrates were mapped into many prevalent diseases. Figure 1.5 shows the highlights of the kinase-related diseases that are related to the AGC kinase group in the serine/theronine kinase category [30].

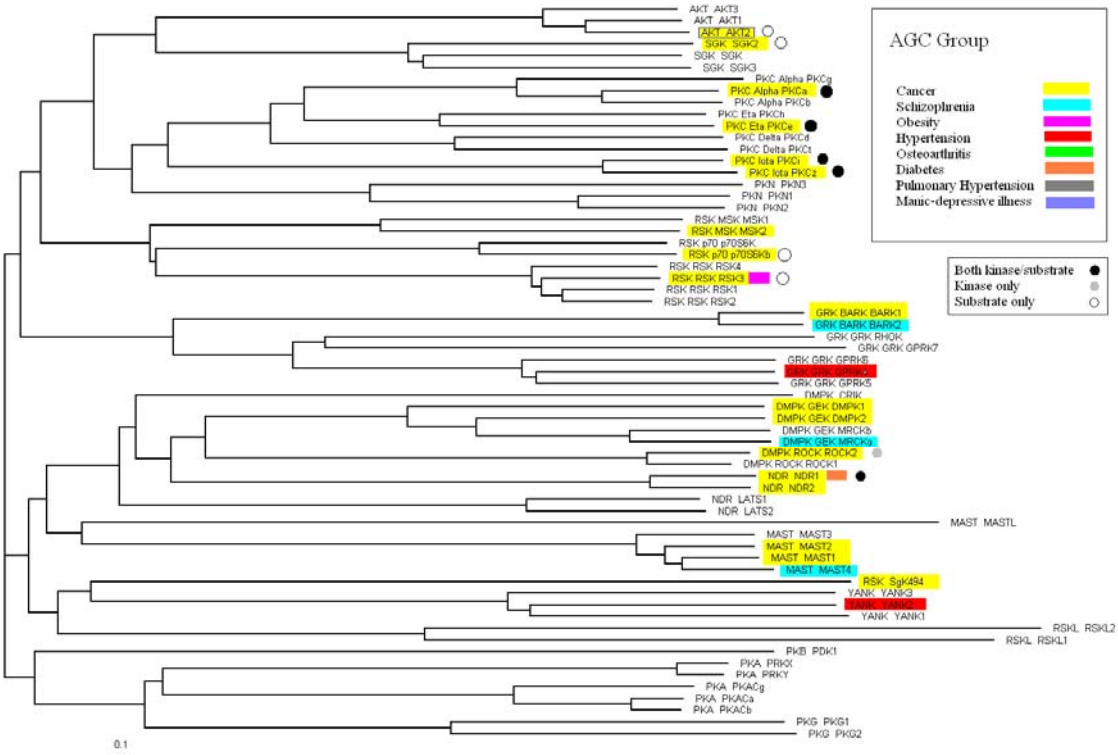


Figure 1.5 Kinase related diseases in AGC group

Eight prevalent diseases were mapped to AGC kinase group for the kinase and its corresponding substrates.

Another candidate group of PPIs that is related to the transcription factor binding activities are selected for analysis. The transcription factors are the

proteins that bind to specific DNA sequences to control the transcription processes between DNA and RNA [31-32]. The transcription factors perform their functions alone or interact with other proteins to form protein complexes [33]. The formation of the protein complex can lead to promoting or blocking the RNA polymerase recruitment [34]. The transcription factors are associated with many diseases/disorders such as cancer [35-36], diabetes [37-38] and the neuro-developmental disorders [39]. The transcription factors are very important in drug development. About 10% of the currently prescribed drugs are directly targeted to the transcription factors [40]. Unlike kinase/substrate proteins in the phosphorylation events, which contain protein binding domains, the transcription factor proteins contain the DNA-binding domains (DBD), the trans-activating domain (TAD) and the signal sensing domain (SSD) [31]. The TAD contains the binding sites for other proteins such as the transcription co-regulators [41].

Currently, there are many databases that are related to the transcription factors. An example is the Eukaryotic transcription factors, their genomic DNA-binding sites and the DNA-binding profiles database (TRANSFAC) [42]. Unfortunately, the TRANSFAC database only contains the transcription factors and their DNA targets, not the protein binding partners. In this project, PPIs related to the transcription factor activities were selected from the Human Protein Reference Database (HPRD) [28, 43] and are filtered by Gene Ontology (GO) [27].

1.4 Current methods for PPI predictions

It is difficult to verify the transition binding activities experimentally due to the short period of their stages [2]. As a result, some indirect methods are applied to observe the changes in the transition stages. One of them is transition state analog testing. The transition state analogs are stable molecules that bind to enzymes more tightly than substrates [44-45]. The binding could cause different observations in the catalysis results, if the analogs are properly designed. The limitation of the experiments is that the analogs cannot perfectly mimic the transition stage [45]. In addition to the experimental methods, computational methods were applied to test the transition binding activities. Many prediction methods were developed over time, including predictions based on the physico-chemical properties of the protein primary structures [46], Gene Ontology (GO) annotations [47], 3D-structures [48-49], and domain/motif interactions [50-52]. Research in PPIs suggests that the protein networks can be explained by functional unit interactions, and many researchers have tried to explore the signatures that are critical in the interaction networks.

1.4.1 Primary structure and associated information

PPI prediction methods that are based on the protein primary structures are the simplest and the most straightforward applications. Most of the primary structures are highly accurate and are available through the National Center for Biotechnology Information (NCBI) [53].

One of these methods was developed by Bock and his colleagues. The prediction processes were based on the Support Vector Machine (SVM), which generates a representation of non-linear mapping from residue sequence to protein fold space [46]. In other words, the representations of each protein interaction (called feature vectors) were applied to the non-linear kernel to differentiate the properties of the interactions. The feature vectors of SVM contain the residue properties such as the charge, the hydrophobicity and the surface extension parameters. The control group was generated by the random selections from the DIP database [54] since the databases of the non-interacting proteins are not readily available [46]. The machine learning process of the data is 2-fold, which means the ratio of the training data and the testing data is 1:1. The results from the SVM process are binary, which indicates the two proteins either bind to each other or they do not.

The accuracy of this method is still considerably high among PPI prediction methods after nine years. The usage of the physicochemical properties provides useful information in the prediction processes because the false positives are removed. However, the requirements for physicochemical information make the method difficult to extend as a global approach. The number of PPIs is small in the prediction (around 4,000). The major limitation of the method is that the predictions cannot be made if the PPIs do not contain their corresponding physicochemical information.

1.4.2 Gene Ontology (GO)

Scientists have tried to develop computational models that do not only depend on the protein sequence information. Wu and her colleagues developed a model solely based on the Gene Ontology (GO) annotations for the yeast protein-protein interactions [47].

The essential algorithm of the method is that the model measures the similarities between the two Gene Ontology (GO) terms with a relative specificity semantic relation (RSS) score. Prior to the real application, the GO terms were filtered if the biological processes or the cell locations were unknown. Since the GO database is constructed as a connected graph, the pathways from the GO term i to the GO term j are not unique for many cases. The algorithm chooses the shortest pathway and separates them into the three cases. The RSS is measured by three parameters: α, β and γ. The α parameter measures the specificity of MRCA (most recent common ancestor) for the two GO terms, which are according to the graphic structures. The β parameter measures the relative generality of the GO terms. The generality of a term is defined as the minimum distance between the terms and all of the leaf terms from it. The γ parameter measures the local distances between two terms relative to the MRCA. Based on the three parameters, the RSS value for the given GO terms i and j are defined from the formula in the paper.

The protein pairs were divided into three categories based on the RSS value: high confidence group, medium confidence group and low confidence group. The authors then constructed the networks that were based on the gold standard. The

positive data consists of 40,753 interactions among 2,259 proteins [47]. The evaluations were performed from the mapping in the MIP complexes database and 35% of the complexes were identified as interconnected.

1.4.3 Geometric simulations for binding interfaces

The proteins interacted with each other with the 3D structures during real biological processes. The large experimental 3D structure databases provided potential possibilities to characterize the binding interfaces. Bahadur and his colleagues generated a model that relies on crystal-packing interfaces in terms of size, shape and packing density [49].

The crystallized structures were extracted from the PDB database (Protein Database Bank) [21]. Two parameters that are related to the interfaces were considered: the size and the shape. The size was calculated as the sum of SASA (solvent accessible surface area) for the different subunits, minus the SASA of the complexes. In addition to the size of the interface, the shape of the interface is another significant feature vector. The term “planarity” is used as a measure of the flatness for the interface curves. “Circularity” refers to the calculation of the ratio of the lengths in the principal axes of the least squares plane that penetrates the atoms at the interface. A circularity score equal to one means the interface is close to a circular shape. “Shape complementarity” is the measure for the interfacial packing of the protein complexes [49]. For the prediction of actual real processes, the chemical compositions of the PPI interface and water molecules were also taken into account.

The major advantage of 3D methods is that the geometric simulation processes exclude many false positives. These false positives could not be removed from the linear protein sequence models. However, the trade-off is that the coverage is limited by the number of protein 3D crystallized structures. Another issue is protein 3D structures are dynamically changed in the biological processes [21]. Existing 3D structures do not guarantee that the structures in actual protein binding process maintain the functions.

1.4.4 Functional signatures

3D crystallized structures require large amounts of time and labor, and therefore some scientists have tried to explore "grammar" embedding in the primary/secondary/tertiary structures of proteins. Domains usually are defined as a conserved region whose functions are involved in the protein interaction process. Motifs are very short residue sites, but play an important role in mediating or regulating protein interactions. In 2005, Albercht and his colleagues decomposed the protein networks into domain-domain interactions (DDIs) and showed that protein interaction processes can be demonstrated by functional signatures embedded in the protein primary structure [55].

Chang and his colleagues developed a model that focuses on the Cyclin-dependent kinase Cdc28 (Cdk1) in yeast [56]. The paper defined the Cdk motif regular expression. The essential method algorithm revolves around is clustering the yeast proteins based on the number of occurrences of motif sites. The clusters were then compared with randomly generated mock proteome groups. A cut-off

value was generated by the minimized sum of the standard errors, which were calculated as the mean value of all compared ratios between yeast and mock. Proteins above the cut-off were considered as Cdk substrate sites.

Guo and his colleagues developed a computational model for the motif site discovery. Motifs are very short pattern signatures, and are therefore hard to define by experimental observations. The authors aligned the interacted protein sequences and compared them with the established motif databases [57]. The proteins were abandoned if they were not in the same cell components or they did not contain any motifs. To calculate the over-represented motifs, the authors used two different statistical analyses: the one-tailed exact binomial test and the one-tailed Fisher's exact test. The binomial test compares the positive data with the random background based on binomial distribution. Fisher's test uses a hypergeometric distribution to explore rare motifs. Another accomplishment of this paper is the discussions related to negative data. The machine learning process in the evaluations requires negative datasets, but non-interaction protein databases were not available. The paper discusses about four different types of negative datasets: random protein pairs, protein pairs separated by cellular locations, the protein pairs that are unrelated in the biological processes and protein pairs involved in different biological processes that are in a low or median confidence level of same cellular locations.

Functional signatures provide the potential candidate binding sites for the protein interactions. Schelhorn and his colleagues developed an integrative approach to associate the protein interaction regions based on the previously

defined databases. The essential algorithm is to build a probability model that fits the protein interaction networks [51]. The probabilistic method maximizes the expected likelihood of observed PPIs. The parameter θ represents the probability of two regions that interact with each other. It has been defined as the corresponding maximum likelihood estimation parameter in the expectation maximization algorithm. The method provided the results of all the possible combinations of the Pfam and the ELM signatures with high confidence.

Predictions by functional signatures are very popular in the current research field for two reasons: 1) Most of the function signatures do not require a 3D crystallized structure (although the algorithms can use the 3D structure to remove the frequent occurrences), and 2) the functional signatures are defined globally, which means different organisms could be tested in the same model.

1.5 SNP altered signatures

PPI prediction methods are useful in exploring protein mechanisms. Functional signatures play a critical role in interaction processes, which has been illustrated in previous results. Compared to predicting the PPIs phenomenon, it is more interesting to explore the potential consequences for biological networks, if the signatures are altered. The resulting observations can link the sequence alterations to the phenotypes, such as for disease or disorder status.

A single-nucleotide polymorphism (SNP) is a DNA sequence variation for one nucleotide at a certain frequency of the genome [58]. About 90% of human genetic variations are caused by SNPs [59], and these variations can affect

developments of human diseases and responses to pathogens, chemicals, drugs, vaccines, and other agents [60-61]. The results from the HAPMAP project provided the information of genetic variances from different populations; as a result SNPs, are not only considered a key to personalized medicine, but also as an indicator of the essential differences across populations [62]. However, there are 1.42 million SNPs defined from human genome in 2001 [63], and currently, over 10 millions SNPs are defined in the dbSNP database from NCBI [64], with only about 4,000 SNPs mapped to disease/disorder status [65]. In order words, compared to the identification of new SNPs, mapping the known SNPs to their corresponding functions are more effective. Due to the large number of existing SNPs, it is impossible to map all SNPs into their genotypes experimentally. As a result, the computational model mapping between the SNPs and genotypes, especially complex diseases, are required.

Non-synonymous SNPs alter the amino acid in protein sequences; therefore, these SNPs have the potential to change the grammar/signatures that are embedded in the protein sequences (Figure 1.6). The silent/active signature could effect the PPI if the PPI relies on the domains or motifs, which would cause the connections to collapse in biological networks. In other words, the SNP-disease mapping model could provide potential explanations for many diseases related to complex biological pathways.

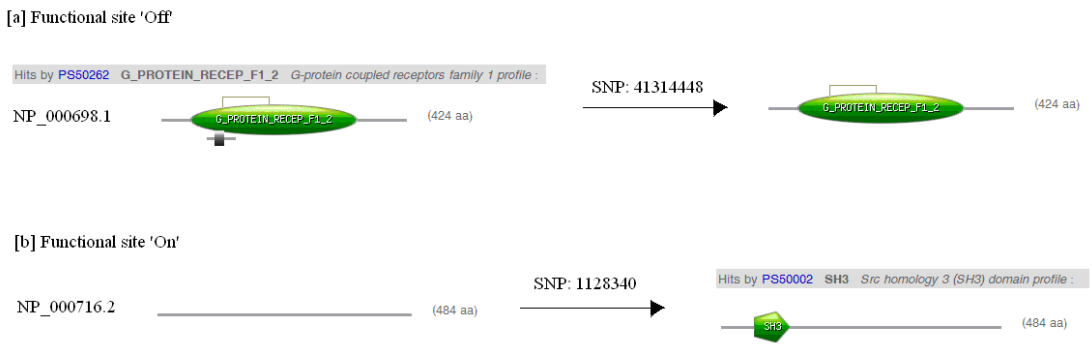


Figure 1.6 Examples of Domain-altering SNPs

Two examples are shown for DA-SNPs. 1.6[a] is an example that the SNP inactive the G-protein receptor domain and 1.6[b] is an example that the SNP active the SH3 domain in the sequence.

1.6 Current methods in SNP-phenotype association

Currently, there are two major categories that are related to SNP-genotype association methods. The first category focuses on non-synonymous SNPs, which change the critical amino acids in a protein sequence, such as binding sites defined in the SWALL database for PolyPhen [66], binding motifs from TRANSFAC for target SNP [67], and the sites related to protein stability and cellular processing in the SNP-effect [68]. These methods are used to select functional SNPs by determining whether the direct alterations present for the critical amino acids (e.g., phosphorylation site). In addition to the false positive problem, the major challenge of these methods is false negatives, which could not explain the fact that certain site changes do not effect genotypes, since the mapping result is a Boolean expression (either modifying the site or not). The second category is based on the computational intelligence model [69]. The SNP-disease data are constructed in different data structures, such as a logic tree (GA, GP) [70-71], neural networks (GPNN) [72], and the ensemble learning approach

(ELAS) [73]. All models use complex data structures to define the SNP-disease association that could achieve good accuracy computationally. The results are also confirmed in clinical data (e.g., Parkinson disease). However, these statistical models emphasize the correlations between the SNPs and the diseases found via observation, which does not provide a potential biological insight in terms of the correlations and/or causations.

1.6.1 Critical amino acid alterations

One group of the SNP-disease association methods focuses on the critical amino acid changes, which affect the binding or structure stabilities. The assumption of the methods is straightforward and the methods mainly focus on the non-synonymous SNPs. These methods (e.g. PolyPhen [66]) use the defined physiochemical databases as dictionaries and check whether the changes for the amino acid would effect the critical sites, such as ligands and other polypeptides. The dictionary for the PolyPhen method is SWALL database [74]. The feature tables in SWALL include DISULFID, THIOLEST, THIOETH bond, BINDING, ACT_SITE, LIPID, METAL or MOD_RES. PolyPhen also maps the substitution sites to the known protein 3D structures. The mapping of an amino acid replacement to a known 3D structure reveals whether the replacement is likely to destroy the hydrophobic core, electrostatic interactions, interactions with ligands, or other important protein features. The application uses the BLAST query to scan the PDB [21] and PQS. A similar approach called SNPeffect [68] defines the SNP-disease associations by protein folding/stability, functional sites and cellular

locations. The folding/stability are estimated by FoldX [[75] and TANGO [76]. The database of functional sites is based on the Catalytic Site Atlas (CSA) [77], which documents enzyme active sites and catalytic residues in enzymes with 3D structures. The cellular localization is determined by PA-subcellular [78] and Psort II [79].

1.6.2 Data-mining/computational intelligence methods

In addition to predictions in the first category, another group of methods that are based on data-mining/computational intelligence was developed in the SNP-disease association. The rationale behind these methods is that the methods should not introduce any human effectors to the system, such as the biological information from external resources. Let the data indicate via different data structures and statistical analysis [69]. Popular examples of statistical methods and machine learning models in SNP-disease study include multiple linear regressions, logistic-regression, Haplotype trend regression, logic regression, random forest, SVM, principal component analysis (PCA) and clustering methods. Both statistical analysis and machine learning models have been tried to find a minimum set of important SNPs that are related to the disease outcomes. However, machine learning models focus more on classification and prediction purpose instead of associate studies in the statistical analysis. Recently developed approaches include GA [70], GP [71], GPNN [72] and ELAS [73].

The GA method categorizes the SNPs into blocks based on the score of the linkage disequilibrium (LD) and constructs logic trees consisting of Boolean

expressions based on these blocks. The logical trees are selected in each generation by using a marginal likelihood in a Bayesian regression framework. The GP method is closely related to the GA method but uses tree-based strategies to represent a solution for the problem instead of using a string of variables. A genetic programming neural network (GPNN) method was developed for detecting epistasis in the SNPs data. Epistasis is the observation that the functions of one gene are modified by one or several other genes [80]. In other words, the GPNN method could detect gene-gene or gene-environment interactions. The GPNN method has been applied to real data analysis of Parkinson's disease [72]. The Ensemble learning approach for the set association (ELAS) method detects a set of loci that predict disease/disorder status. The ELAS method searches the basic learner feature vectors at the beginning and combines the effects of these feature vectors for the SNP prediction. The simulation tests claim that the combination of the markers is more powerful than a single marker. Hubley presented an evolutionary algorithm for SNP selection. The method is based on the modified version of the strength Pareto algorithm, which is particularly suited for multiple objectives involved problems. Other methods related to the evolutionary algorithm were also developed, such as GAs in modeling epistasis and evolutionary trees for haplotype fine mapping [81].

Two major challenges exist for computational intelligence methods. The first challenge is the overfitting problem. Experimental data of SNP-disease associations contain noise due to genotyping errors, missing data and genetic heterogeneity. As a result, overfitting problems are inevitable. For example, GP is

no better than a simple random search when classification accuracy is used as the fitness function. However, if the pre-process was applied with expert knowledge, the GP performance was significantly improved [71]. The problem is that the pre-processing steps violate the rule of the basic principle that no external effectors are allowed to be put into the system. The second potential problem is that the methods provide mappings or associations between the SNPs and the phenotypes, but offer no explanation of the results of the biological process. This fact leads to inconsistency of the simulations and real data. For example, Motsinger applied the ELAS method to large scale type-2 diabetes. ELAS identified 11 significant SNPs and none of them showed significant marginal effects [72].

We merged the designs of both categories, and came up with a bioinformatics approach to predict SNP-disease correlations due to non-synonymous SNPs. The model uses protein sequence grammars as a statistical model and the SNP was selected based on the signal strength that is related to the protein structures or functional domain/motif [18]. The results were then mapped and checked through the human disease database at multiple biological levels, including the SNP level (OMIM) [65], protein level (HPRD disease database) [43], and biological pathway level (KEGG) for evaluations [82]. As a result, the predicted set provided a list of SNP-disease association terms, which provided the explanations. The domain-alter SNPs (DA-SNPs) predicted from the models are the potential targets or markers for further explorations in human diseases.

Chapter 2 introduces the bio-simulation model that constructs domain-domain networks from known phosphorylation events, including the prediction principles and the performance evaluations in the independent databases. Chapter 3 extends the model globally as a software tool and takes the transcription factor PPIs as a test case for evaluations. Chapter 4 introduces the model links of the SNP variations to the phenotypes based on grammar alterations of the functional regions, while the potential effects on biological networks are also explored. Chapter 5 includes the overview of the thesis and two topics for future work that are involved with linkage disequilibrium (LD) and cross-species talk.

CHAPTER 2: MODULAR COMPOSTION PREDICTS KINASE/SUBSTRATE INTERACTIONS.

Note: This chapter is adapted from a paper submitted to BMC Bioinformatics and the requested revisions have been submitted.

2.1 Summary

Phosphorylation events direct the flow of signals and metabolites along cellular protein networks. Current annotations of kinase-substrate binding events are far from complete. In this study, we scanned the entire human protein sequences using the PROSITE domain annotation tool to identify patterns of domain composition in kinases and their substrates. We identified statistically enriched pairs of strings of domains (signature pairs) in kinase-substrate couples presented in the 2006 version of the PTM database. The signature pairs enriched in kinase – substrate binding interactions turned out to be highly specific to kinase subtypes. The resulting list of signature pairs predicted kinase-substrate interactions in a validation dataset not used in learning with high statistical accuracy. The method presented here produces predictions of protein phosphorylation events with high accuracy and coverage. Our method can be used in expanding the currently available drafts of cell signaling pathways and thus will be an important tool in the development of combination drug therapies targeting complex diseases.

2.2 Background

Transient interactions of proteins with other proteins, such as those that occur during phosphorylation events, comprise a fundamental element of signal processing in living cells [14]. Protein kinases constitute one of the largest families of signaling proteins in eukaryotic cells [30]. Currently, there are more than 500 known protein kinases in the human genome [83]. A phosphorylated amino acid distinguishes itself from the unmodified residue by having a large hydrophilic group with increased hydrogen-bonding, hydration and salt-bridge formation capability. Such modifications often result in switches and altered lines of connections in signaling and metabolic pathways of living cells [14]. Phosphorylation binding interactions are important downstream in gene expression pathways in the binding of transcription factors to their substrate proteins [84].

Transient interactions between proteins often require multiple sites of physical connection and may even require a third party protein such as an adopter protein. Catalytic phosphorylation events at active sites is facilitated either by the use of protein recognition modules or the adaptation of docking interactions [85]. Recent structural data indicates that specificity of binding between a kinase and a substrate does not necessarily arise from the active site, but from the substrate and the specific docking interactions [86]. Globular domain – motif interactions accompany active site interactions in the binding of tyrosine kinases to their substrates. Large numbers of such globular domain/linear motif interactions have already been associated with protein-protein interactions (PPIs). Web tools such

as PROSITE [19], Pfam [87], PRINTS [88], ProDom [89], and InterPro [90] can be used to annotate the globular domains and larger linear motifs on the sequence of any given protein. Similarly, the web tool ELM [91] annotates on protein sequences large numbers of linear motifs known to be involved in protein interactions. Some of these motifs may play important roles in virus-host interactions via a mechanism for a hijacking function [92-93].

Known annotations of domain-motif interactions on protein partners often result in the prediction of large numbers of false positives in PPIs [92]. It is also becoming clear that selectivity of docking sites in MAPK kinase, along with the catalytic motif, is an important player in identifying PPIs [94]. An accurate method of PPI prediction based on interactions of short linear motifs on one protein with large globular domains on the protein pair is yet to be developed [95].

Computational prediction of PPIs from primary sequences of proteins poses a number of other challenges to overcome, including the noise in the training PPI data, lack of a true negative training set, as well as problems associated with 3D experimental and molecular modeling of proteins in potentially binding configurations [96]. PPI prediction methods developed in the last decade include methods based on sequence homology [97], feature vectors and machine learning methodology [46], association studies [98], and knowledge guided inference of domain-domain interactions from incomplete PPI networks [99]. Computational studies focusing on extracting domain signature pairs associated with PPIs have utilized yeast datasets [98] or datasets spanning across species [99].

The success achieved in computational association of domain signature pairs with experimentally verified PPIs in these aforementioned studies prompted us to investigate signature pair/PPI associations in phosphorylation events within the human proteome. We asked the question of whether modular composition of proteins (kinase and their substrates), combined with a database of a known PPI, could be sufficient in a statistical enrichment procedure to predict known PPIs not used in the training. The choice of domains as features for predicting PPIs made sense because modular composition of proteins provides insights into their interaction with up and downstream proteins in cell signaling circuits [14, 85].

In addressing this question, we used the Post Translational Modification (PTM) database 2006 edition containing 5,602 PPIs to identify statistically enriched signature pairs in kinase/substrate binding. Our ten-to-one and two-to-one learning and testing procedures produced receiver operator characteristic curves reflecting excellent accuracy in the identification of phosphorylation events. Additional verification included the use of PPIs in the PTM 2009 edition and in other databases not included in PTM [43, 100-101]. Our bioinformatics analysis uncovered sets of domain clusters that are specifically enriched in various kinases and kinase substrates. Moreover, we showed that pairs of such domain clusters bridge kinase and kinase substrates with high specificity and sensitivity.

The computational space in our model is large compared to other approaches focusing only on the domain annotation of proteins known to be interacting with each other. In the present study we scanned the entire proteome for domain annotation in order to develop background sets of randomly generated virtual

protein pairs to be used in the statistical enrichment of domains in protein subsets. Another feature specific to our method is the consideration of strings of domains as signatures for binding predictions. This assumption facilitated us to consider binding events between proteins involving multiple sets of domains. Results produced by our method achieved better PPI prediction accuracy in phosphorylation on average when compared to other presently available computational methods for PPIs. Our study illustrates the dominance of a grammar based on interacting domain signature pairs in the language of post modification interactions between proteins in the human proteome.

2.3 Method

2.3.1 PPI data for phosphorylation events

The learning dataset on kinase/substrate binding was downloaded from the Post Translational Modification database (PTM), version 2006 [28]. The dataset contained 5,602 phosphorylation events between 272 kinases and 1,432 kinase substrates. The independent testing datasets consisted of phosphorylation events not recorded in the PTM 2006 database but recorded in the PTM 2009 database [101], the Human Protein Reference Database (HPRD) [28], and the Biological General Repository for Interaction Datasets (BioGRID) [100]. Predictions of our model were used to match phosphorylated proteins in the PhosphoELM database [91] with candidate targeting kinases for further experimental verification.

2.3.2 Scanning proteins for PROSITE domains and their enrichment in protein subgroups

A database of protein domains, families and functional sites named PROSITE [18] was downloaded to our laboratory's Blade Center. In this set up, the search engine for PROSITE took protein FASTA sequences as inputs and returned hits of PROSITE domains (D) as outputs. Human protein sequences from the NCBI Gene Bank were scanned and a column matrix indicating the presence (1) and absence (0) of domains was assigned for each human protein. The dimension of these domain column matrices was equal to the number of domains (2,102) in the PROSITE Database.

2.3.3 Statistical enrichment of domains in protein subgroups

Statistical enrichment of domains in protein subgroups (target groups) was performed with respect to a control (background) group made of the entire protein kinase group. Domain column matrices were determined for each member of a target group, and these matrices were summed up over the membership of the subgroup. Next, a set of proteins of the same number as the target group was selected randomly from the background group and the corresponding sum domain column matrix was computed. This operation was repeated 10,000 times and the p value for enrichment was computed by the fraction of times the background group had more domains of a given identity than the target group. A list of domains (domain clusters) enriched in a kinase- or substrate subtype was identified as the list of signatures that were enriched in the target group, a kinase- or substrate subtype.

2.3.4 Score matrix for signature pairs in PPIs

A score matrix was constructed for selecting signature pairs strongly associated with known PPIs in tyrosine- and serine/threonine phosphorylation subgroups. Specifically, we wanted to identify signature pairs (such as A-B) such that the presence of signature A in protein K and signature B in protein L would predict with high confidence a PPI between K and L. For this purpose, for the known PPI interactions in the learning dataset (EPPI), we generated a score matrix whose rows and columns identified the enriched signatures in tyrosine and serine/threonine kinases (TK, S/TK) and their substrates (TKB, S/TKB). Each element of the matrix corresponded to the number of EPPIs for which a signature pair (A-B) was present in the opposing proteins of the pair. Another score matrix for virtual PPI and VPPI (background) was generated by randomly pairing proteins from the learning dataset, in effect creating VPPI interactions equal in number to all of the possible protein combinations from the kinase and substrate proteins in the PPI set. The p value for the signature pair enrichment in a given PPI subgroup was computed using the hypergeometric test in the R Project for Statistical Computing, based on the scores summed from the learning set and the background set. The resulting signature pairs were ranked according to their p value, with the one corresponding to the lowest p value ranked highest. The highest p value used as the cut-off in the analysis was $p = 0.001$.

2.3.5 Prediction accuracy for string pairs

The signature pairs thus identified via statistical enrichment were used to predict new PPI events. A protein pair was considered as undergoing phosphorylation interaction if it expressed at least one of the signature pairs determined by the enrichment analysis. Consider a protein pair (L, K) that is associated with a statistically enriched signature pair (A-B). An assumption that the presence of A-B means the presence of a phosphorylation PPI between L and K (PPPI) may lead to false positives. The prediction accuracy was evaluated by computing the probability that the match between predicted and experimental PPI sets have occurred randomly. Consider that there are N VPPI events that can be generated randomly from n kinases and m kinase substrates. Among the N VPPI, M have already been annotated as EPPIs. Let the signature pair A-B predict Y number of PPPIs, W of which have been verified as EPPIs. The hypergeometric test then tests the probability of randomly choosing at least W EPPIs by selecting Y PPPIs out of a possible N VPPI. The lower the p value, the higher the accuracy of the PPI prediction method presented in this study.

2.3.6 Sensitivity, specificity, precision, and recall

In addition to p values, prediction accuracy was evaluated using parameters for defining accuracy and coverage: Specificity (*Sp*) and Sensitivity (*Se*). Let TP, TN, FP, and FN represent, respectively, the true positives, true negatives, false positives, and false negatives determined with the use of known PPIs in the predicted set. *Sp* and *Se* were defined as follows:

Sp = TN/(TN + FP), Se = TP/(TP + FN)

The higher the value of Sp, the lower the error for assuming the PPIs between L and K are based on the presence of the enriched signature pair (A-B). Parameter Se is a measure of the coverage, namely the size of the PPI pool potentially predicted by A-B.

We also used precision and recall to evaluate the statistical enrichment of experimental PPIs in our predicted PPI set. Precision (Pr) was defined as TP/(TP + FP). Recall (Re) is the same as the sensitivity parameter Se.

2.3.7 Cross validation with independent datasets

We used training and testing sets at 2-fold and 10-fold cross validation to test the accuracy of our predictions in 100 iterations using statistical enrichment with p values varying from zero to one [102]. After each set of training and testing, we determined the specificity and sensitivity. We plotted the receiver operating characteristics (ROC) curve using the average values of specificity and sensitivity over 100 iterations. The area under the ROC curve (AUC) quantified the likelihood that one can identify a kinase-substrate interaction using the method described above.

In addition, we used multiple validation processes to evaluate the performance of our model. The first process was to check the accuracy of the enriched signature pairs in predicting PPI among the random protein pairs derived from proteins in the learning data set. A p value representing the probability of

randomly generated prediction was computed for the PPIs predicted by each signature pair by using the hypergeometric test.

Next, we compared the PPI predictions based on the PTM 2006 learning database with PPIs not present in the PTM 2006 database, but present in the PTM 2009 database and in two other databases (HPRD, BioGrid). We identified the phosphorylation PPIs in the HPRD and BioGrid databases as those PPI made of a kinase and a substrate partner of the same type (tyrosine or serine/threonine) listed in Gene Ontology [103]. For each comparison, we computed the number of PPIs predicted, the number of PPIs matched, and the maximum number of virtual PPIs that could be generated using the testing PPI dataset. These numbers yielded p values for random predictions using the hypergeometric test.

2.3.8 Comparison with other computational models

We tested the accuracy of PPI predictions of the present model with two previously published domain based methods: correlated sequence-signature markers (CSSM) [98] and the knowledge-guided inference of domain-domain interactions (K-GIDDI) [99]. Using the algorithms and data presented in these papers, we identified the enriched domain pair signatures and the resulting numbers of predicted PPIs, as well as the number of matched PPIs (matching already annotated PPIs) within the randomly generated PPI set from the PTM 2006 database as well as the validation datasets used for our model. We used the hypergeometric test, sensitivity, and specificity as described above to identify the accuracy of prediction.

2.4 Results

2.4.1 PROSITE domains enriched in kinase and their substrates

Our computations showed that kinases and their substrates express statistically enriched protein domains that are largely subtype specific. We scanned the human protein sequences in the NCBI database via the PROSITE web tool and identified the domains/signatures expressed on their sequences (Figure 2.1 A). We then used statistical enrichment as described in the methods section to identify those domains enriched in a target kinase (substrate) subtype group against all kinase (kinase substrates) with enrichment $p < 0.05$. This enrichment procedure was carried out for the ten kinase subtypes described in Manning's paper. Figure 2.1B shows that domains enriched in a certain kinase (substrate) subtype are largely mutually exclusive to the subtype under consideration. The subtype specificity of domains expressed by kinase and substrates reduced drastically the number of domain signature pairs that needed to be considered for PPI prediction.

Next, we considered the groups of enriched domains expressed by kinases and their substrates, grouped in two major subgroups: tyrosine and serine/threonine kinases (substrates). Many of these proteins expressed more than one subtype-specific enriched domain, as shown in Figure 2.2. In other words, not only domains but domain strings were also enriched in tyrosine and serine/threonine kinase groups and their substrates. Therefore, each such enriched string of domains could be considered to constitute a signature. This observation is consistent with the known preferred mode of interaction between tyrosine kinase

and their substrates (domain-motif interactions) versus the docking site interactions employed in serine/theorine kinases [85].

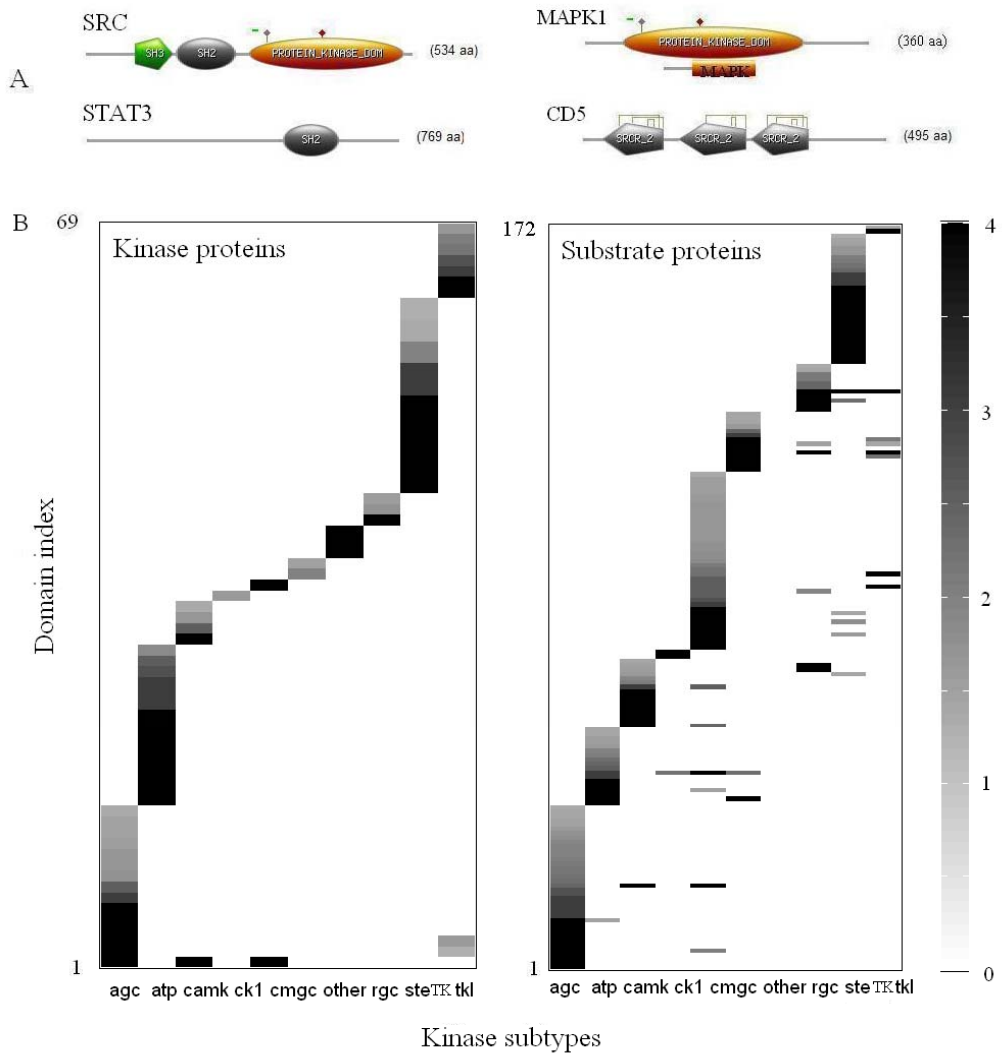


Figure 2.1 Protein domains enriched in kinase- and substrate subtypes

Example of protein domains annotated on the amino acid sequences of kinases and their substrates using PROSITE web tool screen shot (A), their statistical enrichment among kinase and substrate subtypes (B). The horizontal axis in B identifies the kinase (substrate) subtype (agc, atp, camk, ck1, cmgc, other, rgc, ste, TK, tkl) in the notation presented by Manning and others 2002. The vertical axes on the left refer to the identity index of statistically enriched domains in these subgroups of proteins (see Additional File 1 for key to the index). The scale on the right shows the – log p value of statistical enrichment of domains in these protein subgroups.

2.4.2 Score matrices for identifying domain signature sets enriched in known kinase protein interactions

A score matrix in our analysis has m rows and n columns, with each row corresponding to one of the m-enriched signatures (domains or string of domains) in a kinase category (TK or S/TK). Each column indicates one of the n-enriched signatures in the corresponding substrate category (TKB or S/TKB) (Figure 2.2).

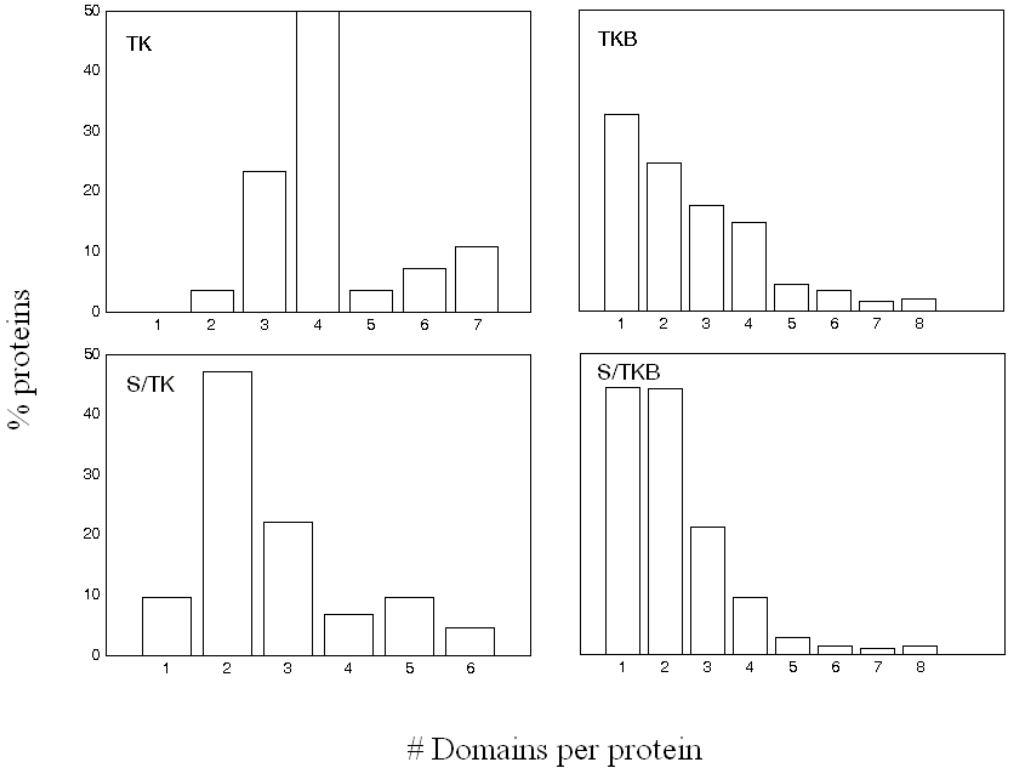


Figure 2.2 Domain number density for kinases and their substrates

Percentage of kinases (substrates) with N domains, N = 1,2, 3, ...8. The symbols TK and S/TK identify tyrosine and serine threonine kinases, respectively. TKB and S/TKB are their substrates.

Elements of the target PPIs score matrix show the number of times a signature pair is found in PPI in the PTM 2006 database. Elements of the virtual PPI score

matrix show similarly the numbers of correlated signatures in this much larger pool of randomly generated protein pairs from the PTM 2006 database proteins in PPIs. Let M be the number of PPIs under consideration and let N be the number of randomly generated protein pairs (including the actual PPI pairs), then a hypergeometric test can be used to estimate the probability of a PPI score matrix element having the value *m* by chance when the corresponding value in a virtual PPI score matrix is *n*. The negative logarithms of these p values for the correlated signature pairs are shown in Figure 2.3 on the score matrix heat maps for TK PPI (left) and S/TK PPI (right). Note that the smaller the p value, the darker the matrix element is corresponding to a signature pair.

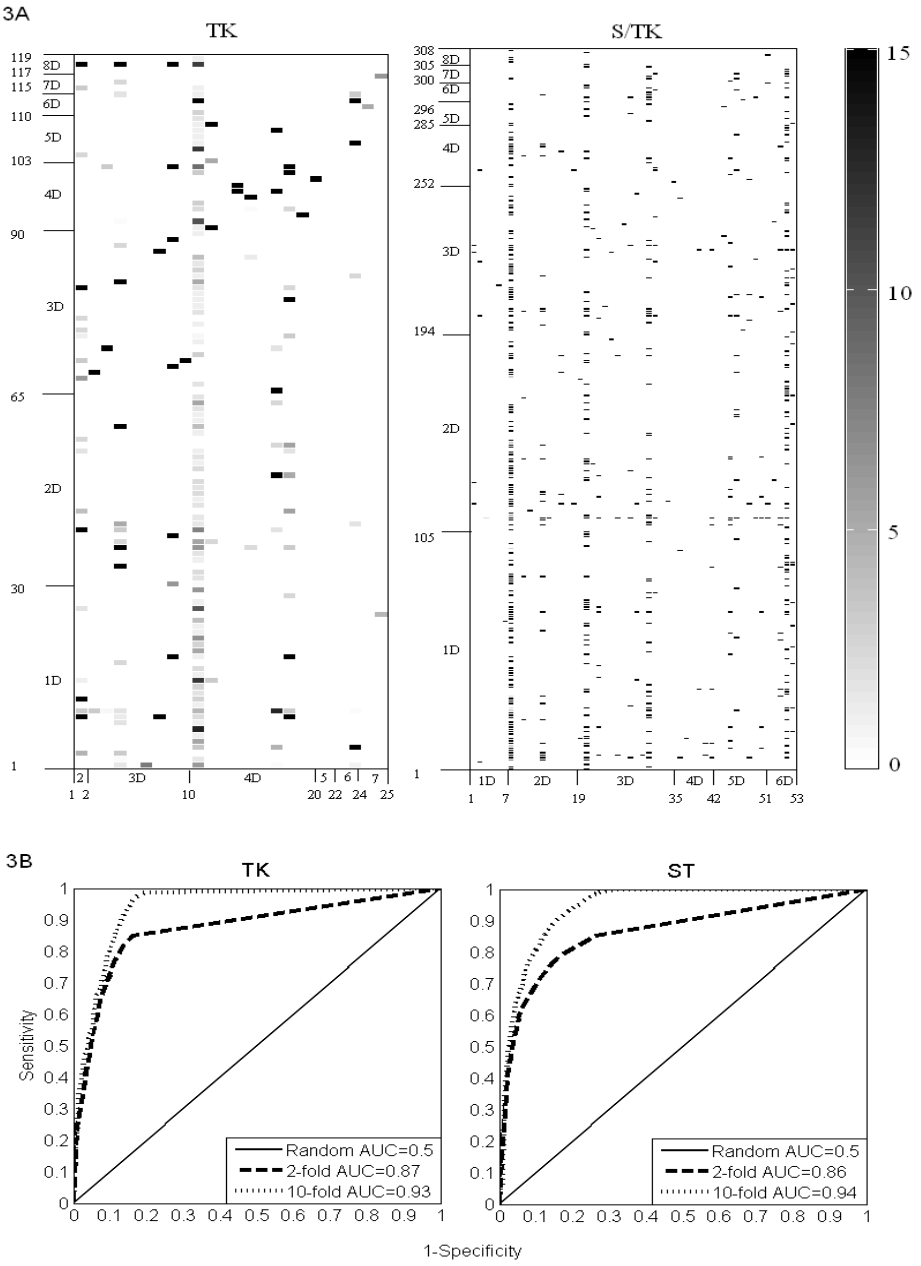


Figure 2.3 Heat maps of the domain signature pairs associated with kinase-substrate interactions & ROC curves in cross-validation

Pairs of strings of domains (one on a kinase, the other on its substrate) that are statistically enriched ($p < 0.001$) in phosphorylation binding interactions in the learning set compared to the set of random pairs of proteins made up of proteins in the same set. The horizontal and vertical axes refer to indices of strings of domains enriched in kinases and kinase substrates, respectively. The scale –log p indicates the level of enrichment of the signature pairs shown in the figure 2.3A. The receiver operating characteristics (ROC) curves for 2-folds and 10-folds cross-validation (2.3B). The area under the ROC curves (AUC) is also shown for the ROC curves in the figure.

The signature pairs presented predicted nearly 80 percent of the PPIs used in identifying the correlated signature pairs. Note that, on average, each signature pair is correlated with ten PPIs, suggesting that domain compositions of proteins involved in phosphorylation are indicative of their potential for binding. The p value shown in Table 2.1 for the training part of this case indicates the efficiency of our score matrix approach in correlating signature pairs with phosphorylation PPI events.

Table 2.1 Accuracy and coverage of the present approach for predicting kinase - substrate interactions. EPPI: Number of Experimental PPI; PPPI: Number of predicted PPIs; MPPI: Number of matches between PPPIs and EPPPIs; VPPIs: Number of Virtual PPI used in p value computations. Also shown is the prediction coverage and accuracy of two previously published approaches (CSSM, K-GIDDI).

		Training		Testing					
		PTM	2006	HPRD		BioGrid		PTM	2009
		TK	S/TK	TK	S/TK	TK	S/TK	TK	S/TK
Data	EPPI	886	2925	137	199	111	166	33	237
	Kinase	56	176	60	104	41	67	15	61
	Substrate	274	881	32	44	26	44	21	90
	VPPI	15344	155056	1920	4576	1066	2948	315	5490
Present	PPPI	1132	7133	43	204	36	69	27	193
	MPPI	617	1876	18	26	17	15	6	16
	p value	0	0	9 E-12	1.3E-7	1E-9	7E-7	0.014	0.0038
	Se	69.6	64.1	13.14	13.07	15.32	0.04	18.18	6.75
	Sp	96.6	95.4	97.76	95.54	96.62	97.66	91.43	96.48
CSSM	PPPI	14496	122975	763	1411	433	836	234	4462
	MPPI	826	2815	84	67	64	53	26	160
	p value	0.9416	0	3.8E-8	0.0096	4.33E-5	0.128	0.204	1
	Se	93.23	96.24	61.31	33.67	57.66	31.93	78.79	67.51
	Sp	5.44	20.69	60.26	69.17	59.38	71.64	25.71	18.72
KGIDDI	PPPI	1491	1557	117	75	88	68	33	65
	MPPI	206	42	19	6	17	5	2	11
	p value	0	0.0098	1.3E-4	0.0432	0.0025	0.181	0.7	0.0002
	Se	23.25	1.44	13.87	3.02	15.32	3.01	6.06	4.64
	Sp	90.28	99	93.9	98.36	91.74	97.69	89.52	98.81

2.4.3 Validation with independent experimental datasets

Approximately 70 percent of known kinase-substrate interactions occurred between proteins with at least one annotated PROSITE domain on their primary sequence. For cross validation, we used the kinase-substrate pair list in the PTM 2006 database and took its subset made of protein couples, with both proteins expressing at least one annotated PROSITE domain. This restriction was necessary since our prediction method is based on the existence of certain domain pairs (signature pairs) in interacting proteins. As described in the methods section, we used 10-fold and 2-fold cross validation in 100 iterations and generated receiver operating characteristic (ROC) curves for predicting tyrosine kinase and serine/threonine kinase interactions (Figure 2.3B). The figure indicates excellent accuracy at 10-fold cross validation and slightly lower accuracy in 2-fold cross validation. The areas under the ROC curves (AUC) for these cases are reported in the figure.

Next, we compared our predicted PPI set with those phosphorylation PPI sets that had not been used in our statistical enrichment processes. Three PPI databases, BioGrid, HPRD, and PTM 2009, contained hundreds of kinase/substrate phosphorylation events, as shown in Table 2.1 for the testing part. We used the signature pairs listed in Appendix B to predict PPI events among the proteins in the PPI events shown in Table 2.1 for the testing set. The p values for the match between our predictions and the known PPI events not used in our enrichment procedures ranged from $9*10^{-12}$ to $7*10^{-7}$ for PPIs presented in the

HPRD and BioGrid databases, whereas we had higher but still significant p values when predicting the PTM 2009 database.

Next, we compared the experimental data shown in Table 2.1 with the corresponding predictions that could be made using the domain based methods recently published (CSSM & K-GIDDI) [98-99]. These comparisons yielded p values that were larger than the ones for our method. In particular, the p values showed no significance for the model CSSM predicting the PTM 2009 database and the serine/threonine binding data from the BioGrid database. The reason why our model yielded better results than CSSM could be due to our grammar differentiating between proteins with different domain string expression. Another reason might be our use of randomly generated background PPI databases in our enrichment method, rather than an analytical equation based only on data for PPIs. Note also that the CSSM model was for the yeast proteome, and we used not their published results, but generated PPI predictions using their procedure here for comparison with experimental data for the human proteome.

K-GIDDI simulation also yielded higher p values than our method when compared with the human protein interactive data shown in Table 2.1. The comparison might be unfavorable to K-GIDDI, since the model incorporates PPI events from multiple species during the training phase and therefore might miss some PPI events specific to humans. Nevertheless, the fact that all three of these approaches gave statistically significant predictions for at least the HPRD database indicates the validity of domain based approaches in predicting phosphorylation events. Sensitivity and specificity parameters were also

computed for the three approaches across different datasets. Present study shows same the accuracy level of specificity as K-GIDDI and better coverage, while CSSM shows much better coverage, but much less specificity.

Overall, our approach predicts 8,837 kinase-substrate interactions from a pool of 186,715 virtual interactions and matches 2,591 PPIs out of the experimentally verified 4,694 PPIs. The p value for the match is zero and precision and recall are equal to 0.293 and 0.552, respectively. Predictions for tyrosine kinase mediated phosphorylation PPIs is better in terms of precision than those PPIs involving serine-theronine kinases (Table 2.2), but, nevertheless, both predictions match experimental data with a zero p value for a random match.

Table 2.2 Efficiency of the present score matrix enrichment in matching known Phosphorylation PPI.

	PPPI	EPPI	MPPI	VPPI	Precision	Recall	p value
Overall	8837	4694	2591	186715	29.3	55.2	0
TK	1238	1167	658	18645	53.2	56.4	0
S/TK	7599	3527	1933	168070	25.4	54.8	0

2.4.4 Matching kinase with substrates in expanding previously annotated cellular pathways

Nearly 30 percent of our predictions match experimentally verified phosphorylation PPIs. We screened the substrates in the remaining 70 percent for their presence in the PhosphoELM [12] database. We found that an additional 30 percent of our predictions involved kinase substrates for which kinase partners are yet to be identified. For this reason, we wanted to see if our PPI prediction

method could be used to revise and possibly expand previously annotated cellular pathways involved in signaling. Consider, for example, the KEGG MAPK signaling pathway [104] showing a chain of phosphorylation events starting at the cell surface and concluding with transcription factors that interact with DNA. A large number of the nodes in the figure are kinase substrates and our DDI based predictions of the corresponding kinases match with those in the KEGG pathway (Figure 2.4). Nodes marked in red in the pathway are listed in PhosphoELM [91] as kinase substrates with unknown kinase identity. Our predicted kinases for those nodes have been added to the KEGG diagram. Out of the 11 predicted kinase/substrate interactions added to the KEGG pathway, 6 appear in the HPRD or BioGrid databases, indicating that any expansions to previously established protein interactomes using our approach will likely be biologically relevant.

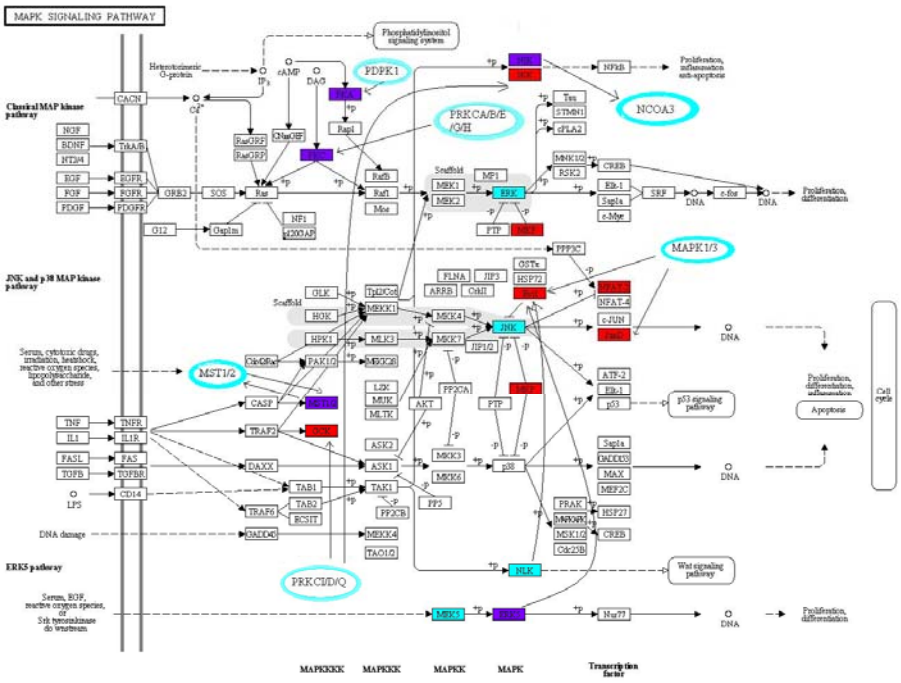


Figure 2.4 KEGG MAPK Pathway revised by adding predicted phosphorylation events
Brackets lined with red identify those nodes in the existing KEGG diagram occupied by a kinase substrate whereas the blue lined brackets identify those occupied by kinases. The proteins that act both as kinases and substrates are shown in purple. The oval shaped nodes are our predictions of kinases that also phosphorylate existing nodes in the KEGG diagram

2.5 Discussion & Conclusion

Binding interactions of proteins with other proteins are at the foundation of cellular networks. Phosphorylation is responsible for the flow of signals and metabolites along the protein pathways [96]. Dynamic binding interactions, such as those that occur in phosphorylation events, appear prominently in signaling pathways in health and in disease, including hypertension, diabetes, HIV infection, and cancer [30, 105]. Although kinases have long been considered as drug targets, compounds targeting kinases (kinase inhibitors and natural substances) have been found to be more promiscuous than originally anticipated, which can potentially lead to side effects [106]. It is important to identify potential phosphorylation partners of kinases in order to assess their range of impact on the flow of signals and metabolites along cellular pathways. Recent methods of mapping dynamic protein interactions in kinase signaling using live-cell fluorescence fluctuation spectroscopy and imaging have already produced promising results [107] and kinase morphisms have been directly linked to population subtypes in disease states [108]. These new experimental approaches will benefit from the ongoing efforts in predicting dynamic protein interactions based on existing data and learning/testing/validation approaches. Our study produces these types of computational prediction sets of protein-protein interactions for experimental validation.

We used large-scale bioinformatics databases and tools and developed a methodology for predicting phosphorylation binding events that are yet to be fully annotated. Our method benefits from the hypothesis and assumptions of the

previous computational methods of PPI prediction and specifically utilizes the concept of correlated sequence signatures as markers of protein-protein interaction developed by Sprinzak and Margalit (2001). The two new elements in our approach consist of (a) expanding the definition of a signature to strings of domains rather than a single domain and (b) the use of a background composed of a random pairing of kinases and substrates in the statistical processes for identifying signature pairs indicative of phosphorylation events. The first assumption is consistent with our observation that certain strings of domains are highly statistically enriched in kinase subtypes and their substrates compared to the rest of the kinase interactome. The second assumption, which was a requirement of statistical enrichment against highly differentiating background sets, allowed us to further reduce the set of correlated sequence signatures obtained solely on the data involving PPIs. The list of signature pairs developed in the present study, when used in predicting kinase/substrate interactions in phosphorylation events, produced results that are largely matched with experimental data not used in statistical enrichments for signature identification. The p values associated with our predictions and their comparison to independent experimental data ranged from a low of E – 11 to 0.0038, depending on the kinase subtype and the database used for comparison.

Thousands of human proteins have been identified as undergoing phosphorylation binding interactions in the PhosphoELM database, but the identity of the kinases responsible for these phosphorylation events have yet to be quantified. Our method produced candidate kinases targeting these substrates. The

resulting list turned out to be consistent with literature not yet included into the PhosphoELM database. In all cases, the partnering between the substrates and the kinases predicted in this study can serve as a guide for kinase identification studies involving known kinase substrates. Another important use of our method will be in expanding and revising existing literature on cellular pathways decorated with phosphorylation events. Such revisions will be useful in identifying the consequences of small drug interventions on a kinase in terms of its interaction with immediate neighbors. Last but not least, our observation that domains expressed by kinase proteins and their substrates are largely subtype-specific drastically reduces the upper bound for the number of experiments one has to conduct for quantifying a major subset of transient binding interactions between protein pairs associated with phosphorylation.

One important disadvantage of our method is the bias toward the discovery of PPIs with proteins having similar domain composition. This feature is also persistent in PPI prediction methods based on sequence homology. This tendency is observable in our prediction of new results included in PTM 2009 based on the PTM 2006 dataset. Although our match is statistically significant, the p values we get for this comparison is significantly larger than the comparison with the HPRD and BioGrid database. It is expected that our methodology will pick up more PPI events correctly as we learn more about the protein sequence grammar that relates domain expression with protein-protein interaction patterns.

Protein phosphorylation events redirect and redistribute the flow of signals and metabolites in cellular pathways. Kinases that phosphorylate multiple substrates

have been favorable targets for drug development against many disease types. In this study, we developed a high throughput method that predicts potential binding partners for kinases using existing domain annotation tools and interactome databases for the human proteome. The method, when tested against independent databases, yields predictions with high statistical accuracy. Results indicate that domains expressed by any two proteins constitute a strong determinant of the potential for phosphorylation related binding interactions between them. Our expansion of the MAPK pathway using the prediction method outlined in the study presented results compatible with research literature.

CHAPTER 3: YiRen: A Prediction Tool for Protein Binding Interactions based on Functional Domain Pair Enrichement

Note: This chapter is adapted from a paper and is waiting for final approval from my advisor for submission

3.1 Summary

Protein-protein interactions (PPIs) play a fundamental role in cell signaling and in response to external factors. Experimentally verified PPI events in the human proteome are like the tip of an iceberg as not a single PPI event has been assigned for more than half of the known human proteins. A computational method predicting candidates for experimental studies on protein binding interactions would be of significant use if the method achieved low false positive discovery ratios. In this study, we introduce a new PPI prediction tool: YiRen. We have developed a tool package to predict potential PPI events based on primary sequences of proteins annotated for their functional domains. YiRen takes a user defined binary PPI set as a training set, generates a domain string pair enriched model called score matrix, then uses statistical enrichment analysis to extract enriched domain string pairs compared with a random background. YiRen has been tested in the PPI ontology related to phosphorylation and transcription factor activities and shows better performance across similar technologies in both PPI prediction accuracy and domain-domain interaction coverage.

3.2 Background

Protein-protein interactions (PPI) play a fundamental role in biological processes spanning from signal transduction, and post translational modification to the assembly of ribosomes and other protein machinery [109]. In the absence of a high throughput experimental methodology for quantifying protein interactions, computational methods have been developed to see if a sequence and 3D structure of proteins could be used to predict the binding interactions between them [96, 110]. These methods, some of them using machine learning techniques, present the opportunity to develop candidate interaction sets for experimental validation.

Interactions between proteins often require multiple sites of physical connection and may even require a third party protein such as an adopter protein. Catalytic phosphorylation at active sites is facilitated either by the utilization of protein recognition modules or the adaptation of docking interactions [85]. Web tools such as PROSITE [111], Pfam [112], PRINTS [88], ProDom [89], and InterPro [26] can be used to annotate the globular domains and large linear motifs on the sequence of any given protein. Similarly, the web tool ELM [24] annotates on protein sequences large numbers of linear motifs known to be involved in protein interactions. The expression of proteins as a series of domains and motifs allows for elucidation of not only binary but also multi protein interactions.

The potential for predicting protein binding interactions based on domain functional units has been under investigation [113]. This research has already demonstrated that the protein networks could be at least partially explained by domain/motif interactions enriched in protein binding interaction subtypes. This

study presents a software tool package, called “YiRen”, for the prediction of protein binding interactions based on the statistical enrichment of domain pairs in binding interaction subtypes. The software uses a set of user defined PPIs as a training set, extracting the domain feature vectors from a global domain database. The code allows for the determination of overrepresented patterns of domain combinations using a “score matrix” approach (Chapter 2). The user could use these patterns to predict new PPIs from their own pool of proteins of interest. The previous research indicated the overflow of false positives in protein binding interaction predictions [92]. The size of the set of PPI predictions presented by the tool depends on the level of accuracy specified and thus could be used to provide predictions with high specificity and low sensitivity, if so desired by the user. The code the tool is based on was used by us previously for predicting kinase/substrate interactions based on domain enrichment (Chapter 2), but it is presented in this study as a bioinformatics tool useful to predict transient interactions between proteins and creating protein networks from a specified group of proteins.

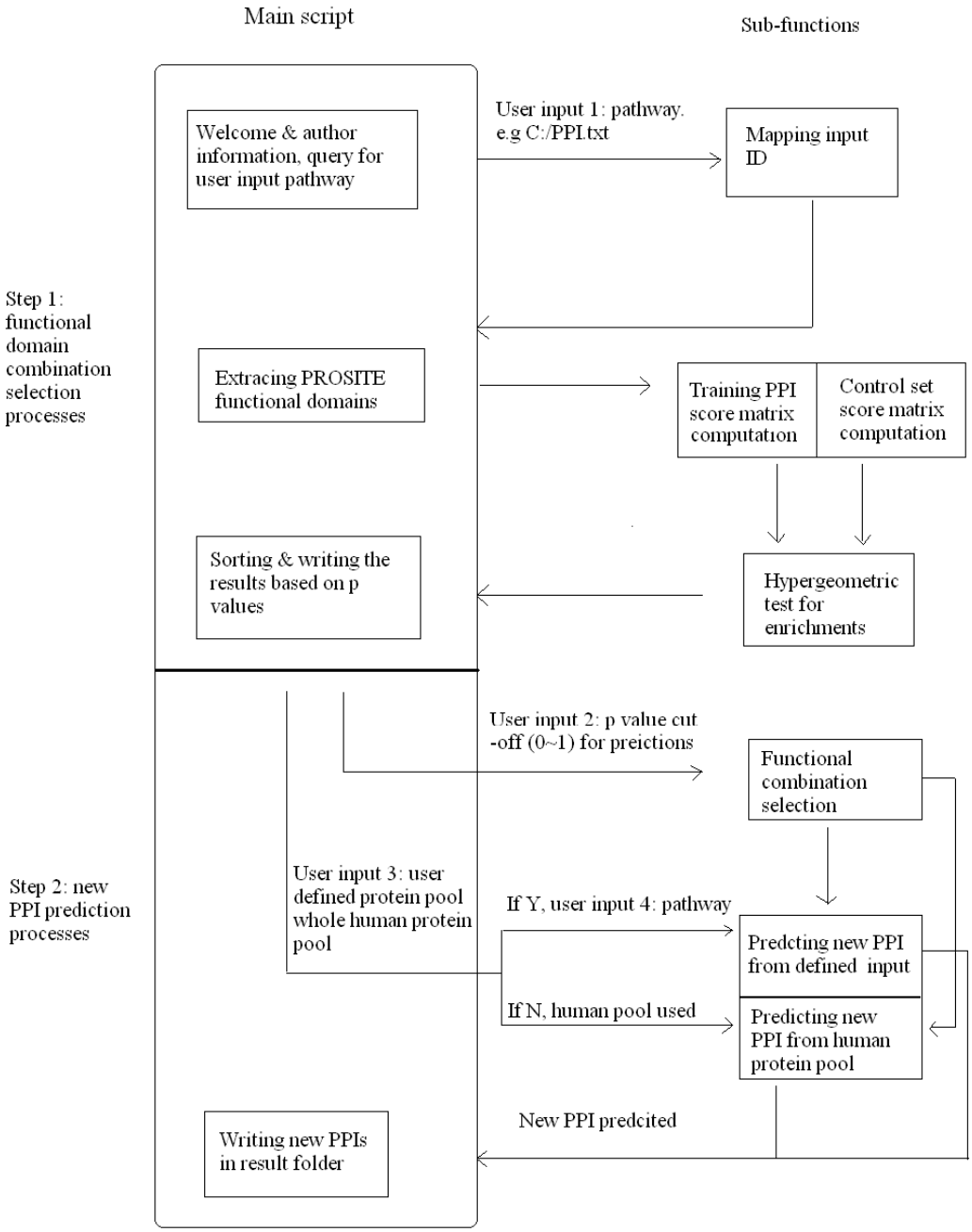


Figure 3.1: Flow chart of YiRen PPI prediction tool

The figure shows the flow direction of YiRen prediction tool. The process is separated into two steps: feature vector extracting process and predicting process in main script.

3.3 Implementations

The flow chart shown in Figure 3.1 summarizes the process of predicting PPIs using YiRen. The algorithm produced for this tool extracts a PROSITE domain expression from protein sequences, generates score matrices for training and testing against random backgrounds, conducts statistical enrichment analysis and predicts sets of new PPI events. To conduct the tasks shown in Figure 3.1, YiRen has been organized as a main script calling sub functional processes in a step-by-step procedure.

YiRen is activated for use when the user double clicks on the "script.exe" file after downloading the tool package from the website http://code.google.com/p/tozerenlab/downloads/list and unzipping it. A command window will show up and welcome information will appear. Then the query for the first user input appears. The user is required to type in the absolute pathway of PPI interaction files (human protein RefSeq ID required). Each PPI in the input file must be according to the format "proteinA + /tab + proteinB" (an example could be found in the "testCase" folder).

After a pathway is correctly typed in as an input by the potential user, the tool maps the input proteins into the human protein database from the NCBI GenBank and extracts the PROSITE domains. A score matrix is generated for selecting domain signature string pairs strongly associated with the training PPIs. The tool identifies signature pairs (such as A-B) such that presence of signature A in protein K and signature B in protein L would predict with high confidence a PPI between K and L. A signature indicates a PROSITE domain or domain cluster.

The method described in detail in Chapter 2 is summarized as follows. For each training PPI interaction in the user input file, YiRen generates a score matrix whose rows and columns identify the enriched signatures. Each element of the matrix corresponds to the number of PPIs in the training set in which a signature pair (A-B) is present in opposing proteins of the pair. Another score matrix for virtual PPIs (control group) is generated by randomly pairing proteins from the training dataset. The p value for enrichment in a given PPI is computed using a hypergeometric test based on the scores summed from the training set and the background (control) set. The resulting signature pairs are ranked according to their p value, with the one corresponding to the lowest p value ranked highest. For an input file containing two thousand PPI events, the score matrix computation and domain string pair selection takes about 10 seconds on a desktop computer. A file name "pV_sorted" is generated in the folder for results. This file includes the domain strings sorted by a hypergeometric p value, a cumulative number of PPI coverage, sensitivity, specificity against the training set and the name (PROSTIE ID) of the corresponding domain strings.

The decision rule used in YiRen for whether a protein pair is a PPI or not is as follows: if the protein pair contains at least one significant domain string pair for a given p value cutoff, it is considered as a PPI; otherwise, it is not. The second user input is in response to the query for the p value cut-off for prediction. After a float number (from 0~1) is typed for the p value, all of the domain string pairs having p values less than or equal to the cut-off are used to annotate new PPI events.

The third user input is in response to the query about the pool proteins among which new PPI events are to be predicted. If the user types 'Y', the fourth user input dialog appears asking for the pathway of the protein list (in RefSeq ID). If the user types 'N', the program takes whole human proteins as candidate pool for prediction. Because of the large number of a human protein pool (~37000) and its possible combinations (>1 billion protein pairs), this process will take hours to finish. A new set of predicted PPIs (in RefSeq ID) is generated based on the decision rule described above and a file name "prediction_PPI" will be located in the result folder. The command window will be close after all computations are completed.

3.4 Results & Discussion

The potential use of the tool is illustrated with protein binding predictions related to transcription factor activity. Prediction accuracy of the tool is compared with two other computational approaches also utilizing domain expression in the prediction of protein binding interactions. Transcription factor binding to other proteins initiates, enhances or inhibits the transcription biological processes [8]. In eukaryotes, many transcription factors do not bind to DNA directly, but interact with the RNA polymerase [9], resulting in multi-protein interactions.

We used a subset of PPI related to transcription factor binding activity in the human protein reference database (HPRD) 2006 version as the input training dataset. We projected all PPI in HPRD on Gene Ontology Molecular Function level 4 transcription factor activity and took the subset of PPIs for which one

member of the PPIs belonged to this GO category. The PPIs related to the same GO category in the HPRD 2009 version (but not in the 2006 version) were applied as a validation set. The validation results were cross-compared with two previously published domain based methods: correlated sequence-signature markers (CSSM) [98] and the knowledge-guided inference of domain-domain interactions (K-GIDDI) [99]. Results are shown in Table 3.1.

Table 3.1: Validation of YiRen prediction across independent PPI set with other methods for 37 PPIs in the new 2009 version from 54 proteins pool.

	YiRen	**CSSM**	**k-GIDDI**
# predicted PPI	59	434	36
# Matches	7	18	2
p value	6.84E-05	0.0043	0.0605

The p values for the match between our predictions and the known PPI events not used in our enrichment procedures is significant in predicting transcription factor activities in the HPRD 2009 version. The comparisons yielded higher p values for the two other models in the literature. In particular, the p values showed no significance (p value>0.05) for the model K-GIDDI predicting the HPRD 2009.

The comparison might be unfavorable to K-GIDDI since the model incorporates PPI events from multiple species during the training phase and therefore might miss some PPI events specific to humans. Nevertheless, the fact that all three approaches gave statistically significant predictions for at least the HPRD database indicates the validity of domain based approaches in predicting events related to transcription binding activity.

In conclusion, the tool package YiRen can be used to identify candidate links among a given set of proteins creating a protein network based on the knowledge of domain pairings in known cases of protein binding interactions. Specific applications to phosphorylation (Chapter 2) and transcription binding activity indicates that predicted links between proteins are highly statistically enriched with links identified and verified experimentally. Our method is applicable not only to the discovery of protein binding pairs but also to the association of multiple proteins in a transient or stationary complex.

3.5 Availability

Project name: YiRen
Project home page: http://code.google.com/p/tozerenlab/downloads/list,
Operating system(s): Windows
Programming language: Python
Other requirements: Python 2.5
License: none
Restrictions of use by non-academics: contact authors for conditions

CHAPTER 4: DOMAIN ALTERING SNPS IN THE HUMAN PROTEOME AND THEIR IMPACT ON SIGNALING PATHWAYS

Note: This chapter is adapted from a paper submitted to Biophysical Journal and is currently under review.

4.1 Summary

Single nucleotide polymorphisms (SNPs) constitute an important mode of genetic variations observed in the human genome. A small fraction of SNPs, about four thousand out of the ten million, has been associated with genetic disorders and complex diseases. The present study focuses on SNPs that fall on protein domains, 3D structures that facilitate connectivity of proteins in cell signaling and metabolic pathways. We scanned the human proteome using the PROSITE web tool and identified proteins with SNP containing domains. We showed that SNPs that fall on protein domains are highly statistically enriched among SNPs linked to hereditary disorders and complex diseases. Proteins whose domains are dramatically altered by the presence of an SNP are even more likely to be present among proteins linked to hereditary disorders. Proteins with domain-altering SNPs comprise highly connected nodes in cellular pathways such as the focal adhesion, the axon guidance pathway and the autoimmune disease pathways. Statistical enrichment of domain/motif signatures in interacting protein pairs indicates extensive loss of connectivity of cell signaling pathways due to domain-altering SNPs, potentially leading to hereditary disorders.

4.2 Background

Hereditary disorders are often linked to rare mutations in the form of single nucleotide polymorphisms (SNPs) [114]. Evolutionary forces introduce many new variants into the human genome in each generation [58]. SNPs affect the tendency to develop autism, diabetes, and cancer and impact immune response to pathogens, chemicals, drugs, and vaccines [59-61]. The HAPMAP project presents information concerning genetic variances among ethnic population subtypes thus implicating SNPs as key differences across population subtypes [62].

More than ten million SNPs have been identified in the human genome [64]. SNPs that fall into coding or promoter regions of proteins comprise only a small fraction of the presently annotated SNPs. To date, nearly four thousand SNPs have been mapped to the disease/disorder status [65]. Genome-wide computational studies complement clinical studies correlating SNPs to disease. However, approaches based on statistics alone provide limited insights on how a genetic variation causes disease.

Current methods for discovery of SNP-genotype linkages include those focusing on non-synonymous SNPs that alter functional motifs such as binding sites [66], DNA binding motifs [67] and sites related to protein stability and cellular processing [68]. Computational intelligence models [69] utilize logic tree [70-71], neural networks [72], an ensemble learning approach [72] or evolutionary algorithms [73] to discover correlations between SNPs and hereditary disorders and provide potential biological insight for the observed correlation and/or

causation. Overall, the aforementioned approaches have illustrated the potential use of computational system modeling in the discovery of links between disease and the genotype.

This study focuses on a specific subset of human genotype-disease linkage, namely the annotation of SNPs that alter protein domains and thus potentially break bonds between interacting proteins in cell signaling pathways [14]. Protein domain structure is relatively flexible with respect to the amino acid sequence defining the domain, as illustrated by the domain annotation web tools such as Pfam [115] and PROSITE [18]. However, scanning proteins through these web tools, one can illustrate that even a single SNP could alter the structural configuration so extensively as to erase a domain from the structural composition of a given protein.

In this study, we screened the human proteome for domain annotation using the PROSITE web tool [19]. We projected the previously annotated SNPs onto proteins and identified those SNPs with domain altering properties. The resulting set turned out to be highly statistically enriched among proteins linked to genetic disorders [18]. We annotated these proteins using a variety of bioinformatics databases and web tools and showed that proteins with domain altering SNPs crowd the protein networks involving focal adhesion, axon guidance, natural killer cell mediated cytotoxity, and neurotrophin signaling pathways. Our predictions of linkages broken in these pathways indicate severe reduction of connectivity in signaling pathways associated with complex diseases and hereditary disorders.

4.3 Method

4.3.1 Discovery of domain-altering SNPs

The human SNP database was downloaded from the NCBI dbSNP database [64]. dbSNP is an archive containing over 10 million human SNPs, and among them, 63,899 missense SNPs. The SNPs were then projected onto corresponding human protein sequences from the NCBI GenBank. Peptide sequences of potentially SNP-containing proteins (in SNP-absent and SNP-containing forms) were screened for annotation of protein domains using PROSITE [18]. A domain with potential to contain a SNP was called D-SNPs.

The PROSITE output for each sequence was in the form of a matrix with three columns, with the columns indicating (1) the ID number of the PROSITE domain, (2) the binary value identifying the presence or absence of a domain, and (3) the PROSITE matching score (MS) if the domain was expressed as a profile (position weight matrix) rather than expressed in the form of a pattern (regular expression). The second and the third columns of the output data allowed us to identify those SNPs that either removed a domain from the protein structure or drastically altered it when compared to sequences with and without the SNPs (DA-SNPs). If the PROSITE domain was defined as a regular expression, the second column was sufficient to identify whether the domain also existed in the presence of the SNP.

For those domains expressed as a profile, we checked the third column for the value of the matching score parameter of the same protein with and without the SNP. We defined a domain distortion (DD) parameter as the ratio of the

difference in the matching score (due to the presence of the SNP) to the matching score of the sequence without the SNP. In our scans, DD varied from 0 to 0.3, which was the maximum domain distortion observed in our computations. Domains with DA-SNPs were defined as the sets of domains for which the sequence with SNPs no longer fits the regular expression, plus a set of profile domains with a finite DD value cut off in the presence of the SNPs.

Examples of structural diagrams of proteins with DA-SNPs were obtained using the Protein Data Bank (PDB) [116] (for the case of no SNP) and SNPs3D [117] (with SNPs). The structures were aligned using YASARA [118] and the location of the SNP was marked with yellow. The lists of proteins with D-SNPs and DA-SNPs were presented as inputs to DAVID Bioinformatics resources [119] and enriched KEGG pathway [120] profiles and Gene Ontology [103] categories at a p-value cut-off of 0.01.

4.3.2 Bonds broken between a protein with a domain altering SNP and its neighbors in signaling pathways

We used statistical enrichment to identify protein signatures (domains, motifs) most likely to be found among binding partners of the proteins containing domain-altering SNPs. We created a score matrix with rows indicating domains that can be altered by an SNP and columns indicating domains and motifs found in binding partners of proteins with domain altering SNPs. Each element of the score matrix represented the number of times a domain with an SNP was found associated with a signature (domain, motif) on a binding partner. The web tool ELM [23] was used to annotate linear motifs on proteins. We then created random

protein binding partners to proteins with SNP containing domains and created a score matrix as a background for statistical enrichment analysis. We used the hypergeometric test to identify those domains/motifs most likely to signal a protein-protein interaction involving domains with DA-SNPs. This procedure allowed us to identify signature pairs (such as A-B) such that the presence of signature A (domain with an SNP) in protein K and signature B in protein L would predict a binding interaction between K and L. The link A-B is a candidate for a bond potentially broken due to the presence of the domain altering SNP in a cellular pathway. To eliminate possible false positives in the estimates of bonds broken, we required the signature pair (A-B) to be either in the DOMINE [121] database or previously annotated as a domain-motif pair as predictive of binding interactions between two proteins [92].

4.4 Results

Our computations show that proteins with SNPs in one or more of its domains are significantly more likely to be associated with human disorders. Out of the 63,899 SNPs in the coding regions of proteins, 1,782 SNPs are present in the Online Mendelian Inheritance in Man (OMIM) database [65]. A total of 12,965 SNPs fall into protein domains, and 592 proteins with domain SNPs are associated with a disease or disorder in OMIM. If this intersection to have occurred by a random event, it has a zero p value, indicating that SNPs in the domain regions of proteins are highly correlated to genetic disorders and complex diseases. This observation

is consistent with the important functions protein domains play in establishing connectivity among proteins in cell signaling pathways [14].

Among proteins with domain SNPs, those with domain-altering SNPs are even more likely to be associated with disorder/disease. Domain-altering SNPs discovered in our PROSITE screening method consists of two subsets. The first subset consists of SNPs the sequence no longer satisfies in the regular expression for the domain. The second subset is composed of domains defined as a profile above a prescribed domain distortion (DD) parameter cutoff. An example of a domain with DA-SNP is p53. Figure 4.1 shows the 3D structure of TP53 in the presence and absence of a DA-SNP (SNP rs28934571), as well as the poor alignment of these structures due to the presence of the SNP. This SNP occurs at sequence position 249 and causes losses of hydrogen bonds and salt bridge bonds.

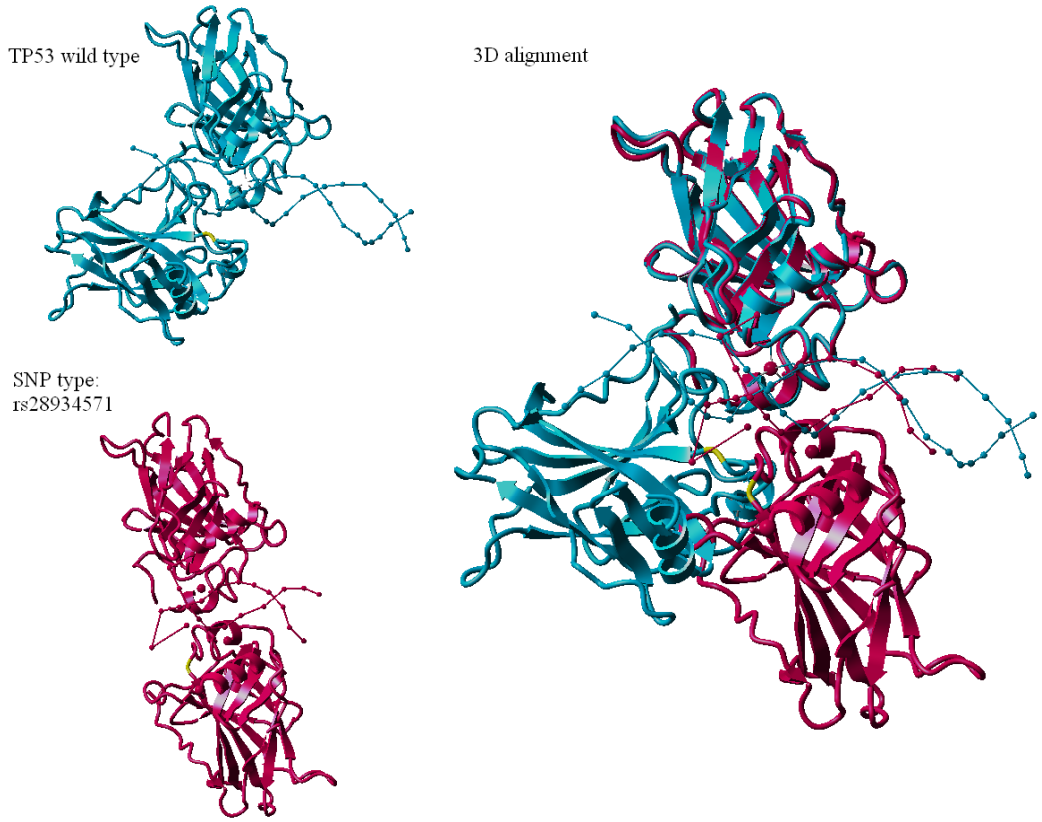


Figure 4.1: Alteration of 3D structure of TP53 due to presence of SNP rs28934571

The TP53 protein wild type (light blue) and the same protein with the SNP (pink) are shown on the left, and their optimal alignment on the right. The SNP position is colored as yellow in resulting structure.

We evaluated the statistical enrichment of the domain altering SNPs in the OMIM Database as a function of the DD cutoff, as shown in Table 4.1. In all cases, the p value < 0.05 indicates significance of enrichment with respect to all domains with SNPs. The list of proteins with domain-altering SNPs in which both the protein and the SNP were linked to the same disease/disorder is presented in Table 4.2 for DD > 0.10. The table covers proteins associated with a variety of diseases ranging from pancreatic cancer, epilepsy, and to carpal tunnel syndrome.

Proteins with domain-altering SNPs crowd GO molecular function categories involving calcium ion binding, adenyl ribonucleotide binding, protein kinase

activity, and endopeptidase activity at DD > 0.10. The DD cutoff of 0.10 corresponds to 598 domain-altering SNPs present in 505 proteins. Among these proteins, 242 had at least one known binding partner in the Human Protein Reference Database (HPRD) [43]. The GO level 5 molecular function gene ontology categories shown in Figure 4.2A are statistically enriched with proteins with domain altering SNPs ($p < 0.01$). The list of GO categories shown in the figure indicates that proteins with domain-altering SNPs comprise key nodes in protein networks; their loss of connectivity would likely have a significant effect on cellular signal transduction.

Table 4.1: Statistical enrichment of domain altering SNPs in the OMIM database.

Each row gives the overall statistics of domain-altering SNPs and the p value for statistical enrichment in OMIM at a given domain distortion (DD) parameter cutoff. The p values are computed based on hypergeometric test.

# D-SNP	# D-SNP & OMIM	DD cut-off	# SNPs	OMIM match	p value
12965	801	0.05	1152	75	0.0444
		0.10	598	46	0.0197
		0.15	497	40	0.016
		0.20	451	35	0.028

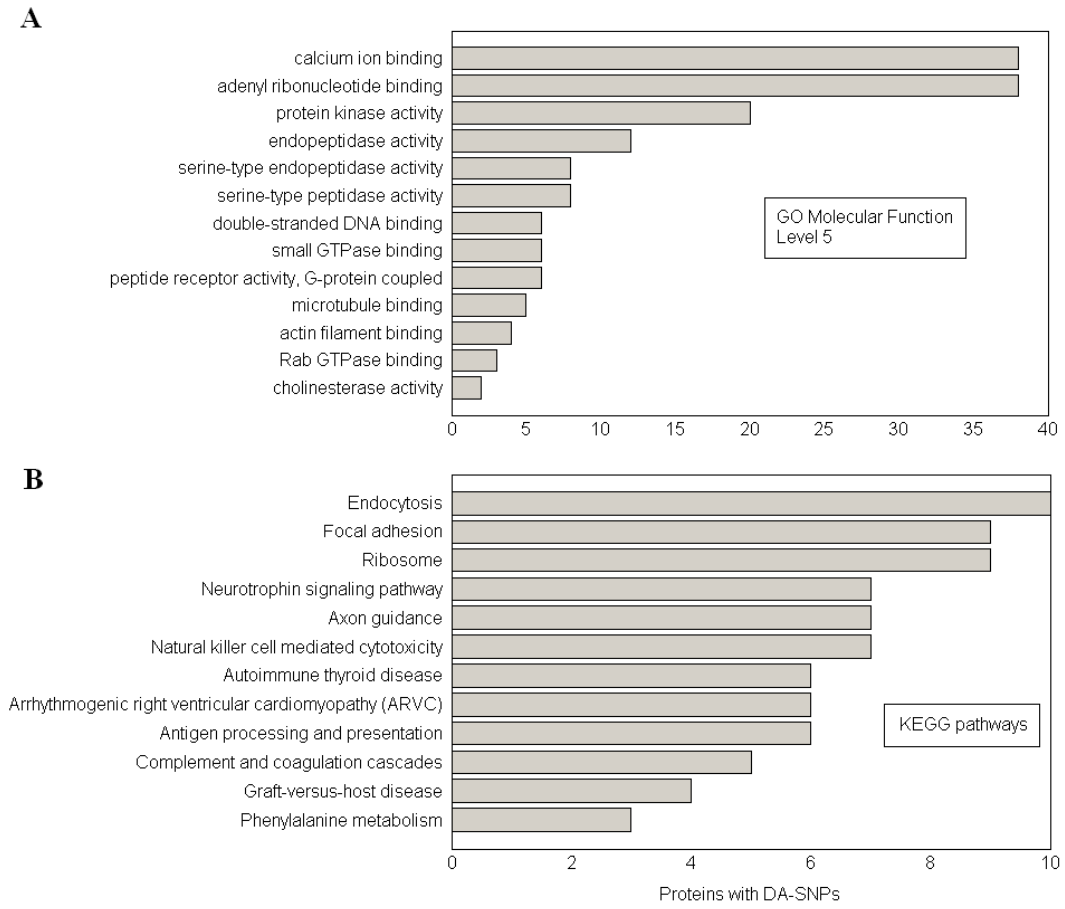


Figure 4.2

Statistically enriched Gene Ontology (GO) molecular function level 5 (MF) categories (4.2A) and KEGG cellular pathways (4.2B) for DD > 0.10 at p value < 0.01.

Table 4.2: List of proteins with domain altering SNPs (DD > 0.1) for which both the SNP and the protein were previously associated in the literature with a hereditary disorder or complex disease. The columns in the table present the gene symbol, the SNP ID, the name of the domain with the SNP, whether is DD violation (yes) or violation of regular expression (no), and the associated disorder/disease.

Gene	SNP ID	PS Domain altered	DD	Diseases
ACADS	28940875	Acyl-CoA dehydrogenases signature 2	No	SCAD deficiency
ATP1A2	28934002	E1-E2 ATPases phosphorylation site	No	Alternating hemiplegia of childhood
CPT2	28936375	Acyltransferases ChoActase / COT / CPT family signature 1	No	Carnitine palmitoyltransferase II deficiency, late-onset

CRELD1	28942091	EGF-like domain profile	Yes	Atrioventricular septal defect, susceptibility to, 2
CRYGC	28931604	Crystallins beta and gamma 'Greek key' motif profile	Yes	Cataract, congenital lamellar
EGFR	28929495	Protein kinases ATP-binding region signature	No	Nonsmall cell lung cancer, resistance to tyrosine kinase inhibitor in
ELANE	28931611	Serine proteases, trypsin family, histidine active site	No	Neutropenia, severe congenital, autosomal dominant 1
F11	28934901	Serine proteases, trypsin family, histidine active site	No	Factor XI deficiency
FBN1	28929500	EGF-like domain profile	Yes	Marfan syndrome, subdiagnostic variant of
FLNA	28935469	Calponin homology domain profile	Yes	Intestinal pseudoobstruction, neuronal, chronic idiopathic, X-linked
GAA	28940868	Glycosyl hydrolases family 31 signature 2	No	Acid alpha-glucosidase, allele 4
GAA	28937909	Glycosyl hydrolases family 31 signature 2	No	Acid alpha-glucosidase, allele 4
KCNH2	28933095	PAS repeat profile	No	Long qt syndrome 2
LGI1	28937874	EAR repeat profile	Yes	Epilepsy, lateral temporal lobe, autosomal dominant
LGI1	28939075	EAR repeat profile	Yes	Epilepsy, lateral temporal lobe, autosomal dominant
MCFD2	28942114	EF-hand calcium-binding domain profile	Yes	Factor V and Factor VIII, combined deficiency of
MCFD2	28942113	EF-hand calcium-binding domain profile	Yes	Factor V and Factor VIII, combined deficiency of
MPI	28928906	Phosphomannose isomerase type I signature 2	No	Congenital disorder of glycosylation, type IB
NKX2-1	28936672	Homeobox domain profile	Yes	Chorea, benign hereditary
NKX2-1	28936672	Homeobox domain signature	No	Chorea, benign hereditary
NOTCH3	28933698	EGF-like domain profile	Yes	Cerebral arteriopathy, autosomal dominant, with subcortical infarcts and leukoencephalopathy
NOTCH3	28933697	EGF-like domain profile	Yes	Cerebral arteriopathy, autosomal dominant, with subcortical infarcts and leukoencephalopathy
NOTCH3	28933696	EGF-like domain profile	Yes	Cerebral arteriopathy, autosomal dominant, with subcortical infarcts and leukoencephalopathy
NOTCH3	28937321	EGF-like domain profile	Yes	Cerebral arteriopathy, autosomal dominant, with subcortical infarcts and leukoencephalopathy
NPR2	28931582	Natriuretic peptides receptors signature	No	Acromesomelic dysplasia, maroteaux type
PAFAH1B1	28936689	Trp-Asp (WD) repeats profile	Yes	Lissencephaly sequence, isolated

RHO	29001653	Visual pigments (opsins) retinal binding site	No	Night blindness, congenital stationary, autosomal dominant 1
RYR1	28933997	MIR domain profile	Yes	Central core disease
SLURP1	28937889	Prokaryotic membrane lipoprotein lipid attachment site profile	No	Mal de Meleda
TMPRSS3	28939084	Serine proteases, trypsin family, serine active site	No	Deafness, congenital neurosensory, autosomal recessive 10
TP53	28934575	p53 family signature	No	Pancreatic cancer
TP53	28934571	p53 family signature	No	Pancreatic cancer
TP53	28934573	p53 family signature	No	Pancreatic cancer
TP53	28934572	p53 family signature	No	Pancreatic cancer
TP53	11540652	p53 family signature	No	Pancreatic cancer
TTR	28933979	Transthyretin signature 1	No	Carpal tunnel syndrome, familial
TYR	28940877	EGF-like domain signature 1; Laminin-type EGF-like (LE) domain signature	No	Tyrosinase polymorphism
WT1	28942089	Zinc finger C2H2 type domain profile	Yes	Mesangial sclerosis, isolated diffuse

Shown in Figure 4.2B is the list of KEGG pathways statistically enriched ($p < 0.01$) in proteins with domain-altering SNPs at DD > 0.10. The list contains pathways closely associated with cancer, neurological, and immunological diseases. Proteins with domain altering SNPs are marked in the pathways for focal adhesion and natural killer cell mediated cytotoxicity in Figure 4.3. Nodes colored in pink in these figures indicate proteins with domain-altering SNPs, while those in blue are their immediate binding partners as identified in HPRD. The purple nodes are proteins that belong to both the pink and blue groups. The pathway diagrams shown in Figure 4.3 illustrate the presence of proteins with domain altering SNPs from the very beginning of the pathway at the cell membrane all the way to the transcription factors regulating important cellular processes.

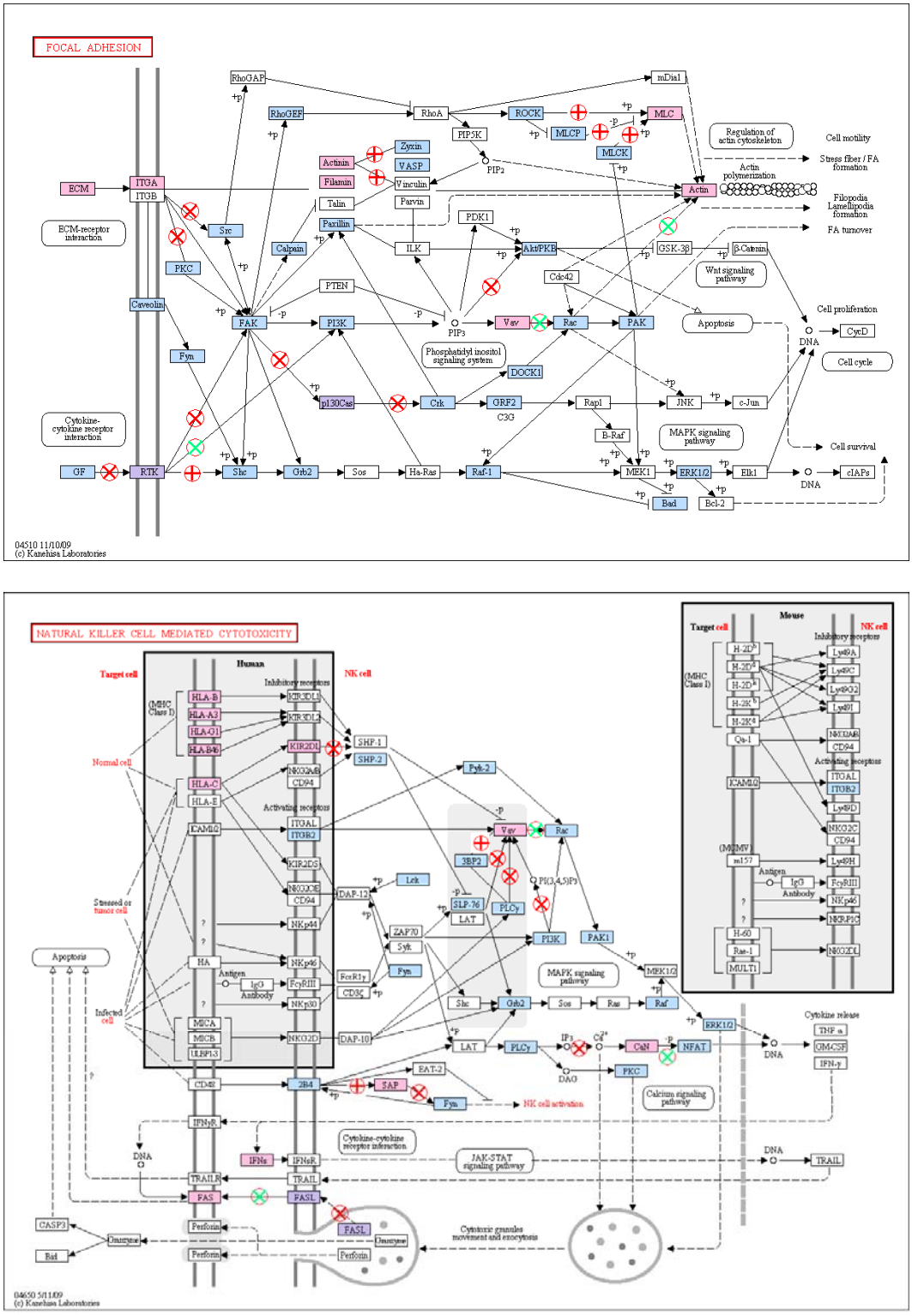


Figure 4.3

Proteins with domain altering SNPs on KEGG pathways for focal adhesion (4.3A) and natural killer cell mediated cytotoxicity (4.3B). Also shown these diagrams are the broken edges (links) estimated by statistical enrichment of domain pairs in protein-protein interactions (circles containing + or x signs.

Next, we estimated the links broken in these pathways due to domain-altering SNPs. We have determined domain-domain and domain-motif pairs (signature pairs) statistically enhanced in protein-protein interactions involving proteins with domain-altering SNPs (Figure 4.4A). The score matrix shown in the figure presents the probability of randomly finding a given signature pair along the sequences of two binding proteins. The darker a point in the score matrix, the higher the association of the signature pair with a protein-protein interaction. In this score matrix, we only considered domains with domain-altering SNPs (total of 60 on the vertical axis) and their possible interacting partners not abundant in the human proteome, meaning they were expressed by less than 25 percent of the human proteins. This resulted in a total of 123 counter signatures, composed of 116 PROSITE domains and 7 ELM motifs on the horizontal axis. The p values for signature pair enrichment were computed by comparing the number of signature pairs observed in binding interactions of proteins with DA-SNPs in known protein-protein interactions (PPIs) with the corresponding number obtained for randomly generated virtual PPIs of the same size. The figure shows that only a very small fraction of possible signature pairs are statistically enriched in PPIs presented in HPRD [26].

We used statistically enriched signature pairs in the estimates of links broken due to a protein expressing a domain-altering SNP. A link (transient or stable) between two proteins is assumed broken due to a domain altering SNP if the opposing protein pair contains at least one signature enriched with the SNP containing domain in the PPIs in HPRD. Shown in Figure 4B is the histogram for

proteins with DA-SNPs, thus indicating the number of edges such proteins have in the absence of an SNP and the number of edges estimated to be broken due to the presence of the SNP. The figure indicates potentially extensive loss of connectivity of proteins with DA-SNPs to neighboring proteins in protein networks.

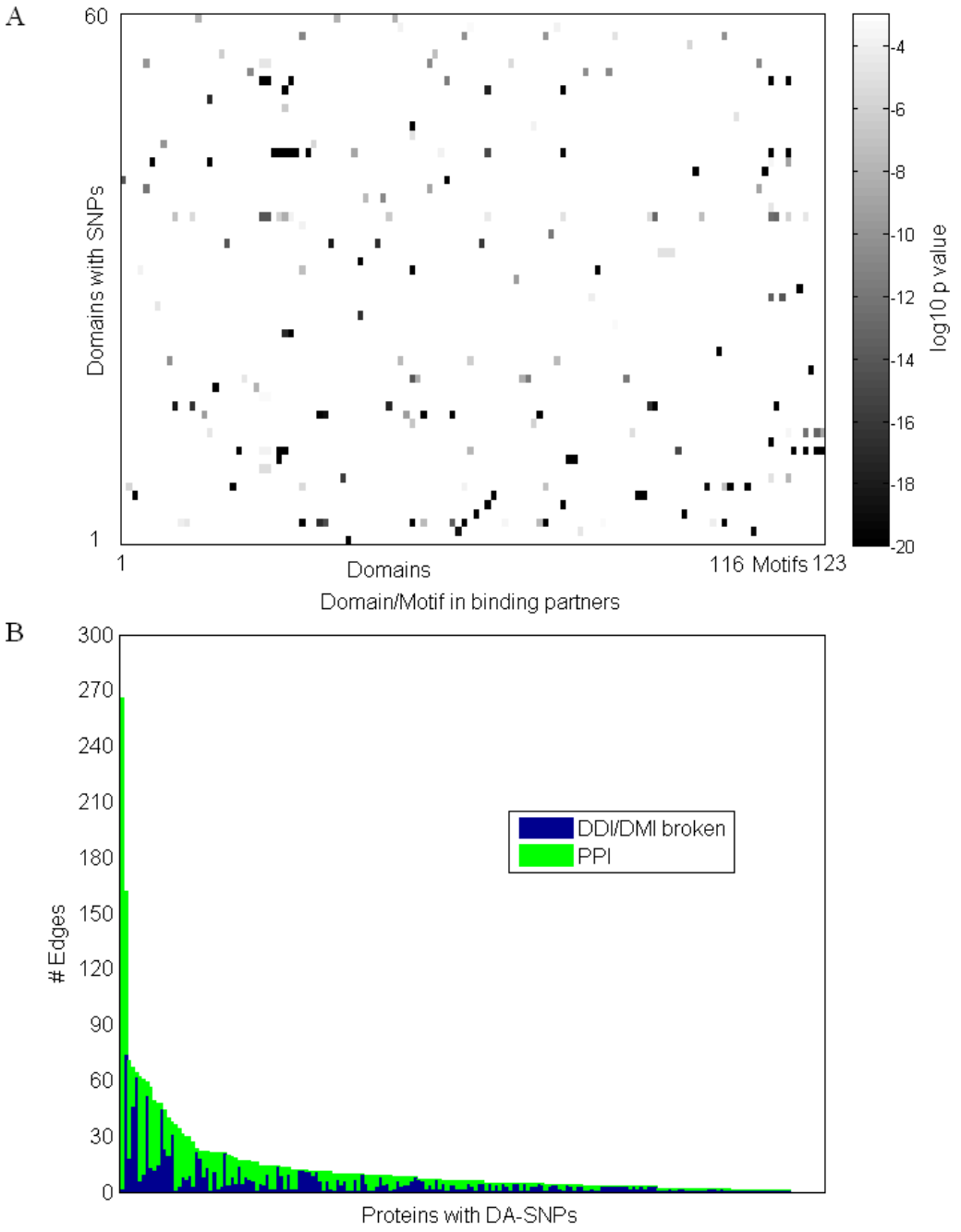


Figure 4.4

Statistical enrichment of signature pairs in binary protein interactions in involving at least one protein with a domain-altering SNP (4.4A). Each point in the score matrix indicates the probability of occurrence p (log10 based) by chance of a signature pair in a protein-protein interaction. Figure 4.4B shows a histogram for proteins containing a domain-altering SNP indicating the number of edges each has in the absence of SNP, and the estimate of broken edges in the presence of the SNP.

The protein links we estimate to be broken between a protein containing a DA-SNP and its immediate neighbors are shown in Figure 4.3A and 4.3B for focal adhesion and natural killer cell mediated immunity. The links deemed to be broken using the statistical enrichment method described above are shown as marked with a circle containing either a red or green "x" or a red "+" sign. The links with red "x" correspond to the domain signature pairs present in the DOMINE database described as predicting PPIs with high accuracy [27]. The links with a green "x" correspond to domain-motif associations deemed highly predictive of PPIs [25]. The links marked (+) are links estimated to be broken by statistical enrichment of the domain pairs only. Note that, even when we exclude these latter links, the pathway connectivity (number of edges per node) is strongly influenced when an SNP drastically alters the 3D shape of a domain responsible for connections to upstream and downstream partners. This assertion is further enhanced by the list of highly connected proteins with domain altering SNPs given in Table 4.3. For example, the transcription factor TCF3 has 44 known binding partners, and links to 19 of these partners could be broken due to the presence of the domain-altering SNP. Taken together, our study exposes the importance of domain SNPs in the progression of some of the most prominent complex and/or hereditary diseases.

Table 4.3: The top ten most highly connected proteins with domain altering SNPs (DD>0.10) and the estimates of the number of broken edges for each protein using statistical enrichment .

Gene	Protein Name	# broken	# intact
ADRBK1	beta-adrenergic receptor kinase 1	29	9
SH2D1A	SH2 domain-containing protein 1A	12	0
TCF3	transcription factor E2-alpha	19	25
APBA2	amyloid beta A4 precursor protein-binding family A	8	1
NCK1	cytoplasmic protein NCK1	61	3
PRKCH	protein kinase C	8	0
YWHAE	14-3-3 protein epsilon	46	21
TOPBP1	topoisomerase (DNA) II binding protein 1	11	3
BCAR1	breast cancer anti-estrogen resistance protein 1	45	3
RIMS1	regulating synaptic membrane exocytosis protein 1	11	3

4.5 Discussion and conclusions

Recent advances in high throughput sequencing facilitated a number of genome-wide studies seeking a correlation between genetic makeup and cardiovascular diseases [122], autism [123], and diabetes [124]. Results implicate a handful of SNPs correlated with these complex disease states. Correlation is based on purely statistical methods, and in many cases, SNPs found to be significantly associated with a disease fell into the non-coding regions of DNA distant from a protein coding gene [69]. As the population subsets for genome-wide studies grow in size with increasing research efforts and time, and as these sets are better controlled for demographic and environmental variables, one would expect the discovery of sets of additional SNPs strongly correlated with hereditary disorders and disease subtypes. Nevertheless, a system bioinformatics approach is needed to explore how a disease-correlated SNP alters cell signaling and metabolic pathways, thus contributing to the initiation of a disorder or a disease.

This study explores the mechanisms by which SNPs that fall into protein domains in the human genome potentially contribute to disease. Protein domains are functional units closely aligned with post-transcriptional modification (as in phosphorylation) and play important roles in establishing the connectivity of cell signaling networks via binding to upstream and downstream proteins [60, 125]. Our computations indicate that the p value for disease association of SNPs that fall into protein domains and occur by random chance is practically zero. Within this group of SNPs, those with domain-altering properties are even more likely to be associated with a disease state. We have defined a domain-altering SNP as one that either alters the sequence such that it no longer satisfies the regular expression of the domain or that the domain is extensively deformed as quantified by the domain distortion index. Proteins with domain-altering SNPs crowd cellular pathways involved in neurological, and immunological diseases, as well as in cancers such as the pancreatic cancer.

How does an SNP with domain altering properties affect the connectivity of pathways? The key to answering this question lies in the discovery of the set of proteins that bind to proteins under consideration via the domain containing the SNP. The grammar of protein-protein interactions in terms of primary sequence and/or 3D structure is yet to be fully understood. We used a statistical enrichment approach to identify protein domains (motifs) on the opposing protein most frequently associated with the SNP-containing domain under consideration. We then assumed a bond (transient or steady) was broken whenever we came across such a signature pair among the immediate partners of the protein with a domain

altering SNP. Results shown in the present study for focal adhesion and the natural killer cell mediated cytotoxicity pathways indicate extensive loss of connectivity in these cellular pathways, caused by the presence of domain altering SNPs among the proteins in these pathways. Even when we reduced the estimates of bonds broken with the use of signature pairs already known to predict protein-protein interactions, the loss of connectivity persisted at multiple cell compartments. We obtained qualitatively similar results for axon guidance and neutrophin signaling pathways altered by the presence of a domain altering SNP (not shown).

In conclusion, proteins with domain-altering SNPs are statistically enriched in the list of proteins known to be associated with disease. These proteins crowd pathways associated with immunological, neurological and cardiomyopathy disorders. Protein functional groups statistically enriched with proteins with domain altering SNPs include calcium ion binding, adenyl ribonucleotide binding, protein kinase activity, endopeptidase activity, serine-type peptidase activity, DNA binding, and GTPase binding proteins

CHAPTER 5: FUTURE WORK

5.1 Overview

Protein-protein interactions (PPIs) are one of the most important categories of signal transitions, because the human protein tends to modify the signals via protein binding, ion binding and post-translational modifications, rather than create new protein sequences [40, 45].

This dissertation has tried to develop a system that explains the signal transduction processes based on built biological functional units. The project tried to build a connection between the feature vectors that are related to the protein sequence and the biological phenotypes. Large sets of known PPIs (currently over 38,000 interactions, which are involved with over 9,000 proteins) [43] provide the available training data for the machine learning process. In the prediction steps, we concentrated on two groups of PPI events, phosphorylation and transcription factor activity, instead of focusing on global PPIs. Both groups are significant in protein signal processes related to transitional binding and the group either contains a specific database (e.g., PTM) or can be filtered by the Gene Ontology database from the global protein interaction database. As a result, the most informational signal or signal combinations were generated through a score matrix model in Chapter 2, and the network of domain-domain interactions (DDIs) or domain-motif interactions (DMIs) can be constructed (Figure 5.1). The reason

for selecting a group of proteins rather than the global protein set is that the binding mechanisms are significantly diverse for different protein types. The protein types have to be related to DDIs or DMIs if the feature vectors that are applied in the system are protein binding domains/motifs.

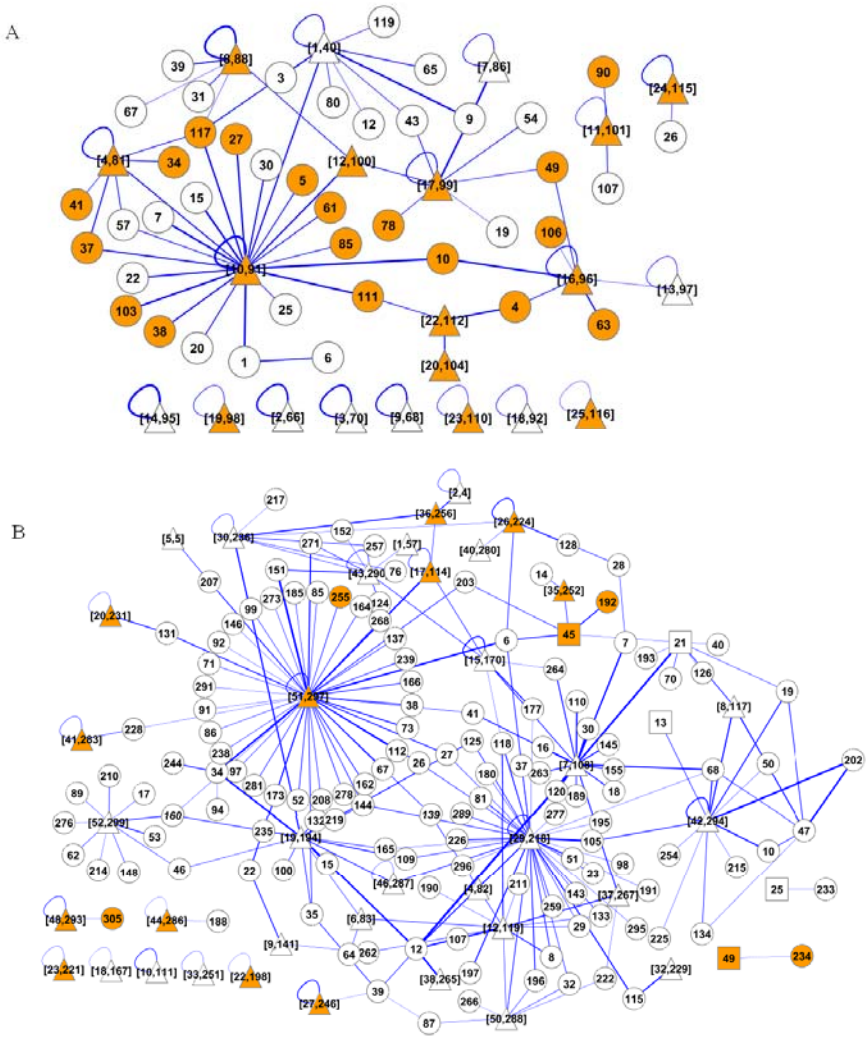


Figure 5.1 Connectivity maps of the DDI interactions

Predicted network for the phosphorylation events mediated by tyrosine (A) and serine/threonine kinases (B). Each node in these graphs represents a domain (square: kinase domain, circle: substrate domain, and triangle: a domain that is enriched in kinase and substrate subtypes). The edges between the nodes identify the domain pairs found enriched in PPI. The thicker the edge the higher is its frequency of occurrence in phosphorylation events. The red color in these figures identifies those domains that are found in the high accuracy DDI subset in the DOMINE database.

After defining the most important signals, the functional strings were cross-validated with independent databases (e.g., PTM 2009 version), 2-fold with random protein pairs, and 10-fold with random protein pairs. The results show excellent prediction abilities with extremely high specificity.

The linkages between the sequence variation parameter (SNPs) to the disease/disorder phenotype were built in chapter 4. The mappings were achieved by evaluating the alteration that is related to domain/motif grammar signal strengths. Currently, there are two categories of methods that are related to SNPs and phenotype association predictions. The methods in the first category take the defined functional sites, such as phosphorylation sites, and create mappings based on the alterations of the critical amino acid. The SNP could disable the unit (e.g., phosphorylation site) if it alters the critical site, which affects the biological networks. Therefore, the SNPs modified in the functional sites are the potential causation/correlation of the diseases. The contribution of these methods is that they provide reasonable chain explanations for the SNP-disease associations. The problem, however, is that is these methods cannot explain the exceptional cases. For example, changes in the critical amino acid do not cause the corresponding diseases. The methods in the second category are based on data mining technology. The assumption is that the input data are full of errors, and the post-filtering processes will introduce additional human errors into the system. As a result, the rationale of the methods is to let the data interpret the phenomenon without adding any additional information into the system. The accuracy of these methods has been proven in many diseases/disorders, such as Parkinson's

diseases. However, the prediction only provides new observations without any potential explanations. Chapter 4 has tried to build an SNP-disease association model based on the grammars embedded in the protein sequences. The model performs as an interface to link the SNPs and phenotypes. As an intermediate interface, the model provides the potential explanation of the phenotypes from the observations that are related to the grammar alterations. The results showed that the disease-related SNPs are highly enriched in the domain regions, and the DA-SNPs have even better enrichment measurements of the diseases. The "inactive" DDIs/DMIs were then projected to the biological networks to identify the broken edges. As a result, the model provided a list of DA-SNPs and the functional signatures to investigate. The altered functional signatures could be potential drug targets in drug development.

5.2 Tag SNPs among different populations

SNPs occur frequently throughout the human genome. Therefore, they are the biomarkers for the human diseases/disorders in the genome-wide association study (GWAS) [126]. The International HapMap Project selected four populations to include Utah residents with ancestry from northern and western Europe (CEU), Han Chinese in Beijing (CHB), Japanese in Tokyo (JPT), and Yoruba in Ibadan (YRI). The HapMap project genotyped the individuals among different populations and proved that the SNPs for different populations are variant [127]. In other words, the variances exist among different populations and the variances are the potential causes of different drug responses among different populations. The distribution of the SNPs (Figure 5.2), which were generated from the

HapMap project raw data, shows that the "global" SNPs only accounted for half of all defined SNPs, while about 30% of the SNPs are shared among two or three different populations.

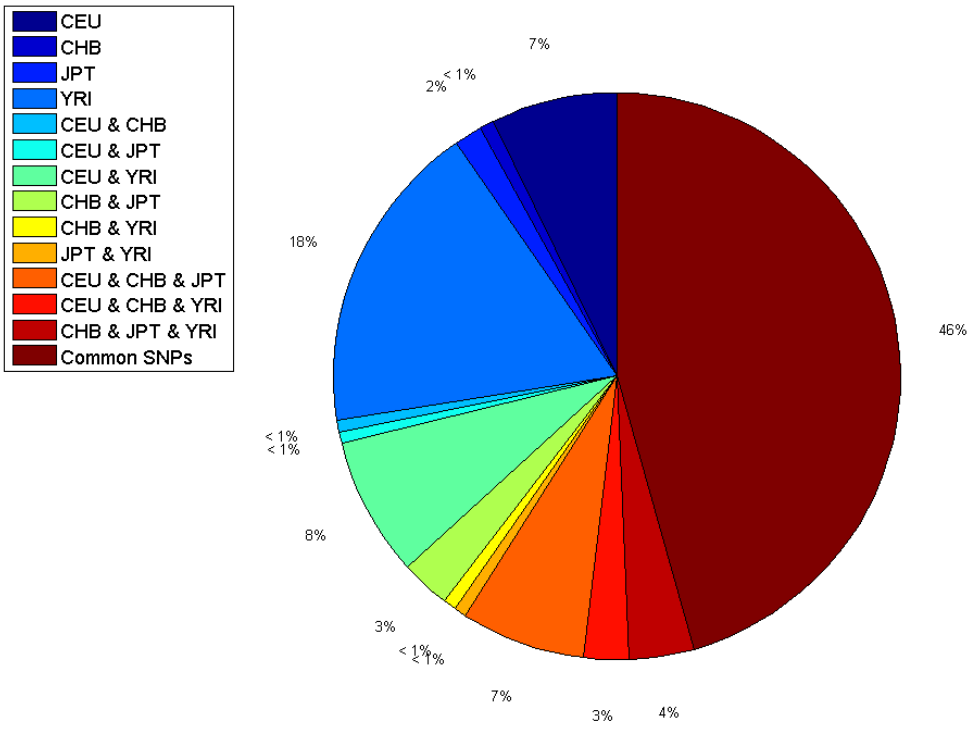


Figure 5.2 Distribution of SNPs among different populations

Four populations were selected in HapMap project, and from its raw data a pie chart represent the distribution of all SNPs cross different populations was shown. Less than half of the SNPs are prevalent among four populations.

Currently, it is too expensive to genotype all available SNPs across the human genome. For diagnostic purposes, a minimum number of SNPs with the most accuracy prediction abilities have been selected, called tagSNPs [69]. The selections of tagSNPs are based on unsupervised methods with the haplotype block concept and the linkage disequilibrium (LD) [128]. The LD describes the tendency of alleles to be coinherited in each generation if the alleles are close to each other on the same chromosome [128]. In other words, the tagSNPs are the

SNPs in the genome with high LD score, which could be coinherited with other SNPs.

TagSNPs are not shared among different populations, either (Figure 5.3). TagSNPs that were defined in the HapMap project were projected into the OMIM database. 64% of the tagSNPs are common among four populations. About 15% of the tagSNPs are unique to one population. Table 5.1 shows some examples of tagSNPs related phenotypes that are uniquely expressed among certain populations.

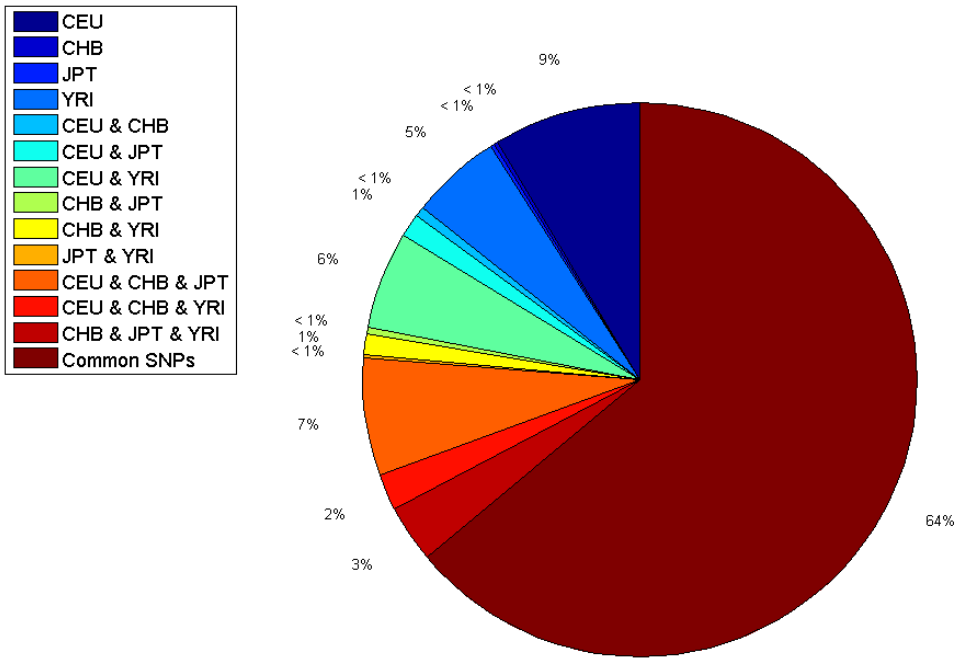


Figure 5.3 Distribution of tagSNPs that are related to diseases/disorders

TagSNPs defined in HapMap were mapped into OMIM SNP-disease association database. 64% of the tagSNPs are in commom and about 15% tagSNPs are unique in one population group. The results indicate the tagSNPs are converged among populations.

Table 5.1 Examples of population-specific tagSNPs and corresponding phenotypes

Population	Diseases/Traits description
CEU	SKIN/HAIR/EYE PIGMENTATION 6, BLOND/BROWN HAIR [SLC24A4, G/T] MENTAL RETARDATION, AUTOSOMAL RECESSIVE 10; MRT10 NEUROPEPTIDE Y POLYMORPHISM [NPY, LEU7PRO] SKIN/HAIR/EYE PIGMENTATION 3, LIGHT/DARK SKIN [TYR, SER192TYR] CELIAC DISEASE, SUSCEPTIBILITY TO, 13; CELIAC13 KIAA1109 GENE; KIAA1109 CELIAC DISEASE, SUSCEPTIBILITY TO, 6; CELIAC6 PHOSPHODIESTERASE 8B; PDE8B APNEA, POSTANESTHETIC, DUE TO BCHE, ATYPICAL-1 [BCHE, ASP70GLY] HYPOADRENOCORTICISM, FAMILIAL OSTEOARTHRITIS SUSCEPTIBILITY 1 [FRZB, ARG200TRP] DIABETES MELLITUS, INSULIN-DEPENDENT; IDDM NEUROPEPTIDE Y; NPY DIABETES MELLITUS, INSULIN-DEPENDENT, SUSCEPTIBILITY TO [PTPN22, -1123, C-G] SCHIZOPHRENIA SUSCEPTIBILITY LOCUS, CHROMOSOME 13q-RELATED DISCS LARGE, DROSOPHILA, HOMOLOG OF, 5; DLG5 LACTASE PERSISTENCE [MCM6, IVS13, C/T] MENTAL RETARDATION, AUTOSOMAL RECESSIVE 7; MRT7 SKIN/HAIR/EYE PIGMENTATION, VARIATION IN, 2; SHEP2 NUCLEOTIDE-BINDING OLIGOMERIZATION DOMAIN PROTEIN 2; NOD2 LUTEINIZING HORMONE POLYMORPHISM [LHB, TRP8ARG AND ILE15THR] MELANOCORTIN 1 RECEPTOR; MC1R SKIN/HAIR/EYE PIGMENTATION, VARIATION IN, 3; SHEP3 G30 GENE OBESITY FAT MASS- AND OBESITY-ASSOCIATED GENE; FTO MELANOMA, CUTANEOUS MALIGNANT, SUSCEPTIBILITY TO, 7; CMM7
CHB	CYCLIN-DEPENDENT KINASE INHIBITOR 2B; CDKN2B DIABETES MELLITUS, NONINSULIN-DEPENDENT; NIDDM ANEURYSM, INTRACRANIAL BERRY, 6 AORTIC ANEURYSM, FAMILIAL ABDOMINAL 3
JPT	Pseudoxanthoma elasticum is an inherited multisystem disorder of the elastic tissue leading to skin disease as well as ocular and cardiovascular complications (Struk et al., 1997).
YRI	HEMOGLOBIN G (MAKASSAR) [HBB, GLU6ALA] PSEUDOXANTHOMA ELASTICUM; PXE ALCOHOL DEPENDENCE [TAS2R16, LYS172ASN] THYROXINE-BINDING GLOBULIN, SLOW [TBG, ASP171ASN] VASCULAR CELL ADHESION MOLECULE 1; VCAM1 HEMOGLOBIN C (GEORGETOWN) [HBB, GLU6VAL AND ASP73ASN] SKIN/HAIR/EYE PIGMENTATION, VARIATION IN, 11; SHEP11 GAMMA-AMINOBUTYRIC ACID RECEPTOR, ALPHA-4; GABRA4 JUXTAPOSED WITH ANOTHER ZINC FINGER GENE 1; JAZF1 HEMOGLOBIN S (CAMEROON) [HBB, GLU6VAL AND GLU90LYS]

	HEMOGLOBIN JAMAICA PLAIN [HBB, GLU6VAL AND LEU68PHE] HEMOGLOBIN S (ANTILLES) [HBB, GLU6VAL AND VAL23ILE] HEMOGLOBIN S (OMAN) [HBB, GLU6VAL AND GLU121LYS] HEMOGLOBIN S (PROVIDENCE) [HBB, GLU6VAL AND LYS82ASX] HEMOGLOBIN S (TRAVIS) [HBB, GLU6VAL AND ALA142VAL] HEPATITIS C VIRUS, RESISTANCE TO [IFNG, -764C-G] TYROSINASE-RELATED PROTEIN 1; TYRP1 GLYCOGEN SYNTHASE KINASE 3-BETA; GSK3B CELIAC DISEASE, SUSCEPTIBILITY TO, 7; CELIAC7

In chapter 4, we proved that the functional signatures (domains/motifs) are highly correlated to the disease-related SNPs. Unique tagSNP markers among different populations indicate that it is possible to develop a functional signature database, not only based on the functions or binding targets of the proteins, but also related to the preferences among different populations. In other words, focusing on the population-specific sequence biomarkers could lead to the development of new domains/motifs for certain population groups.

5.3 Functional signatures in cross-species talk

Protein-protein interactions occur among different species, such as virus-infectious processes. The virus proteins have specific targets of certain species. For example, the HIV-1 virus infects human individuals, but not mice, rats or chickens. However, the sequences are highly similar to each other for different eukaryotic organisms. As a result, exploring potential biomarkers for the virus-specific organisms becomes significant for infection prediction processes. The basic assumption is that the functional signature frequencies for the virus and its corresponding host organisms should be similar to each other, and different when compared to other organisms.

Figure 5.4 shows the frequencies for 81 ligand binding motifs from ELM among all human protein sequences and hepatitis B virus sequences, which are calculated as occurrences per 100 amino acids. The figure shows that although the number of proteins is different in the virus and host, the frequencies are similar related to the ligand binding motifs.

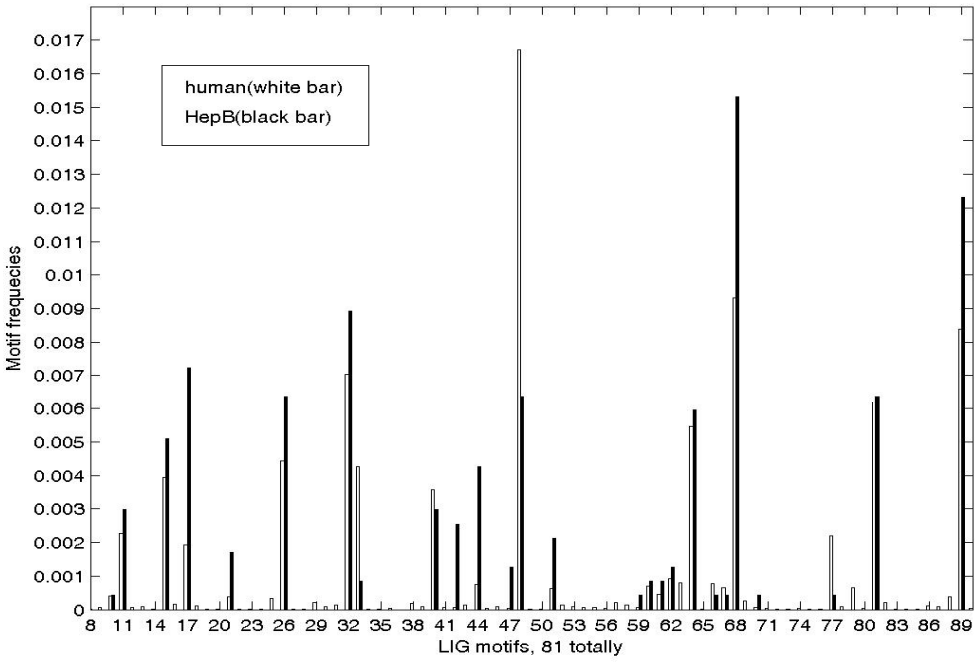


Figure 5.4 Frequencies of 81 ligand binding motifs of human and hepatitis B

The frequency is defined as the motif occurrences per 100 amino acids. White bar (human) and black bar (hepatitis B) have similar frequencies although the number of proteins is different. The ligand binding motifs were extracted from ELM database.

The frequencies of the domains/motifs were then further calculated among five different species, including humans, mice, zebra fish, chickens and rats (Figure 5.5). The functional signatures were combined as one feature vector to compute the angle between the virus and different hosts. An example shown in Table 5.2 indicates the angles of 81 ligand binding motifs among humans and four types of virus (hepatitis A, B, C and HIV-1). The "+" sign indicates the large variation

(motif expression is different). Based on these primary results, it is possible to develop a model to select species-specific biomarkers from functional signatures.

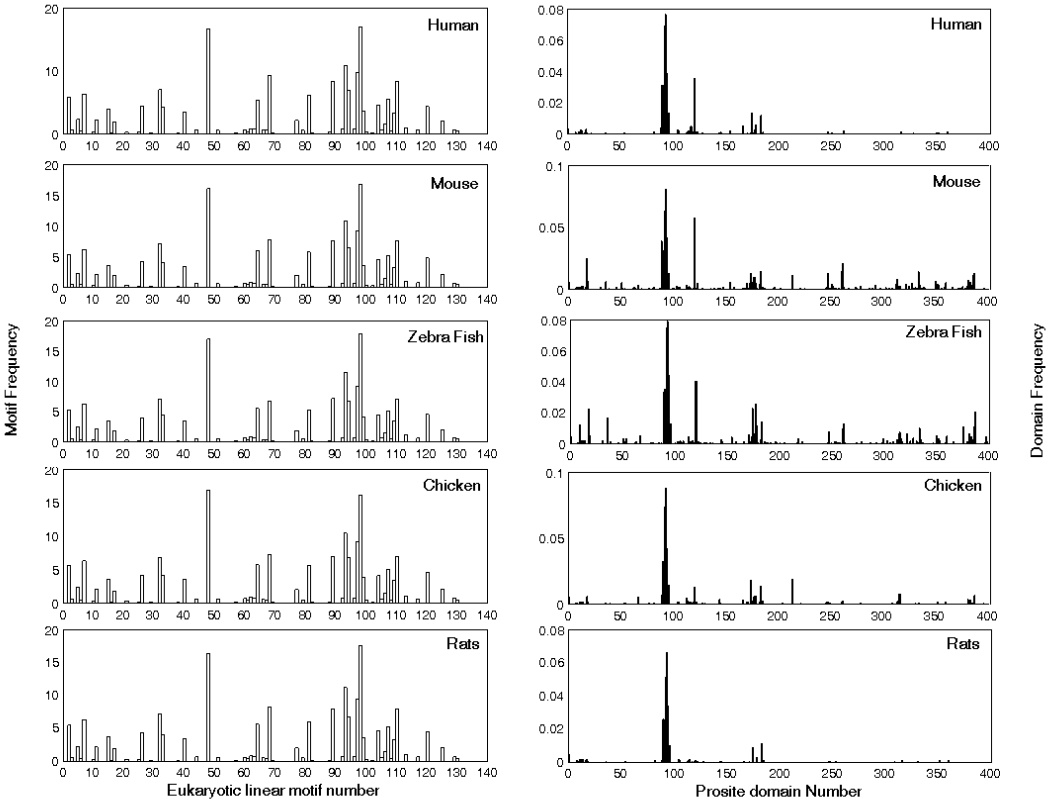


Figure 5.5 Frequencies of PROSITE domain and ELM motifs across species

The frequencies of human, mouse, zebra fish, chicken and rat were computed for 133 ELM motifs and 400 PROSITE domains.

Table 5.2 Angles of ELM ligand binding motif vectors across species

The table contains the angles between species (human, hepatitis A, B, C and HIV-1), calculated based on ligand binding vectors. "+" indicates the significant of motif frequencies.

	HepA_human	HepB_human	HepC_human	HIV_1
LIG_14-3-3_1	0	0	0	0
LIG_14-3-3_2	0	+	0	0
LIG_14-3-3_3	+	+	+	+
LIG_AP2alpha_1	0	0	0	0
LIG_AP2alpha_2	0	0	0	0
LIG_AP_GAE_1	0	0	0	0
LIG_APCC_Dbox_1	+	+	+	+

LIG_APCC_KENbox_2	0	0	0	0
LIG_BRCT_BRCA1_1	+	+	+	+
LIG_BRCT_BRCA1_2	0	0	+	0
LIG_BRCT_MDC1_1	0	0	0	+
LIG_CAP-Gly_1	0	0	0	0
LIG_Clathr_ClatBox_1	0	+	+	+
LIG_Clathr_ClatBox_2	0	0	0	0
LIG_COP1	0	0	0	0
LIG_CORNRBOX	0	0	0	0
LIG_CtBP	+	0	0	0
LIG_CYCLIN_1	+	+	+	+
LIG_Dynein_DLC8_1	0	0	0	0
LIG_EH	0	0	0	0
LIG_EH1	0	0	+	+
LIG_EVH1_I	+	0	0	+
LIG_EVH1_II	0	0	0	0
LIG_FHA_1	+	+	+	+
LIG_FHA_2	+	+	+	+
LIG_GYF	0	0	0	0
LIG_HOMEOBOX	0	0	0	0
LIG_HP1_1	0	0	0	0
LIG_IBS_1	0	0	0	0
LIG_IQ	+	0	0	+
LIG_MAD2	0	0	0	0
LIG_MAPK_1	+	+	+	+
LIG_MAPK_2	+	0	0	+
LIG_MDM2	0	+	+	0
LIG_MYND	0	0	0	0
LIG_NRBOX	+	+	+	+
LIG_PCNA	0	0	0	0
LIG_PDZ_1	0	0	0	0
LIG_PDZ_2	0	+	0	0
LIG_PDZ_3	+	+	+	+
LIG_PIP2_ANTH_1	0	0	0	0
LIG_PIP2_ENTH_1	0	0	0	0
LIG_PP1	+	+	+	+
LIG_PP2B_1	0	0	0	+
LIG_PTAP	0	0	0	+
LIG_PTB_1	0	0	0	0
LIG_PTB_2	0	0	0	0
LIG_PXL	0	0	0	0
LIG_RB	0	0	+	+
LIG_RGD	0	0	+	0
LIG_RRM_PRI_1	0	+	0	0
LIG_SH2_GRB2	+	+	+	0
LIG_SH2_PTP2	0	+	+	+
LIG_SH2_SRC	+	+	+	+
LIG_SH2_STAT3	+	0	+	+

LIG_SH2_STAT5	+	+	+	+
LIG_SH2_STAT6	0	0	0	0
LIG_SH3_1	0	+	+	+
LIG_SH3_2	0	+	+	+
LIG_SH3_3	+	+	+	+
LIG_SH3_4	0	0	+	+
LIG_SH3_5	0	+	0	+
LIG_SIAH_1	0	0	0	0
LIG_Sin3_1	0	0	0	0
LIG_Sin3_2	0	0	0	0
LIG_Sin3_3	0	0	0	0
LIG_TNKBM	0	0	0	0
LIG_TPR	0	0	0	0
LIG_TRAF2_1	+	+	+	+
LIG_TRAF2_2	0	0	0	0
LIG_TRAF6	+	0	+	+
LIG_ULM_U2AF65_1	0	0	0	+
LIG_USP7_1	+	+	+	+
LIG_USP7_2	+	0	+	0
LIG_WH1	0	0	0	0
LIG_WRPW_1	0	0	0	0
LIG_WRPW_2	0	0	0	0
LIG_WW_1	0	0	0	0
LIG_WW_2	0	0	0	+
LIG_WW_3	+	0	+	0
LIG_WW_4	+	+	+	+

LIST OF REFERENCES

1. Solomons TWG: **Organic chemistry**. New York: Wiley; 1976.

2. Lehninger AL, Nelson DL, Cox MM: **Lehninger principles of biochemistry**, 4th edn. New York: W.H. Freeman; 2005.

3. Kraut J: **How do enzymes work?** *Science* 1988, **242**(4878):533-540.

4. Schramm VL: **Enzymatic transition states and transition state analog design**. *Annu Rev Biochem* 1998, **67**:693-720.

5. Johnson LN, Barford D: **The effects of phosphorylation on the structure and function of proteins**. *Annu Rev Biophys Biomol Struct* 1993, **22**:199-232.

6. Capra M, Nuciforo PG, Confalonieri S, Quarto M, Bianchi M, Nebuloni M, Boldorini R, Pallotti F, Viale G, Gishizky ML *et al*: **Frequent alterations in the expression of serine/threonine kinases in human cancers**. *Cancer Res* 2006, **66**(16):8147-8154.

7. Weinberg RA: **Cancer Biology and Therapy: the road ahead**. *Cancer Biol Ther* 2002, **1**(1):3.

8. Lee TI, Young RA: **Transcription of eukaryotic protein-coding genes**. *Annu Rev Genet* 2000, **34**:77-137.

9. Thomas MC, Chiang CM: **The general transcription machinery and general cofactors**. *Crit Rev Biochem Mol Biol* 2006, **41**(3):105-178.

10. Lobe CG: **Transcription factors and mammalian development**. *Curr Top Dev Biol* 1992, **27**:351-383.

11. Ottolenghi C, Uda M, Crisponi L, Omari S, Cao A, Forabosco A, Schlessinger D: **Determination and stability of sex**. *Bioessays* 2007, **29**(1):15-25.

12. Shamovsky I, Nudler E: **New insights into the mechanism of heat shock response activation**. *Cell Mol Life Sci* 2008, **65**(6):855-861.

13. Benizri E, Ginouves A, Berra E: **The magic of the hypoxia-signaling cascade**. *Cell Mol Life Sci* 2008, **65**(7-8):1133-1149.

14. Bhattacharyya RP, Remenyi A, Yeh BJ, Lim WA: **Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits**. *Annu Rev Biochem* 2006, **75**:655-680.

15. Wheelan SJ, Marchler-Bauer A, Bryant SH: **Domain size distributions can predict domain boundaries**. *Bioinformatics* 2000, **16**(7):613-618.

16. Jones S, Stewart M, Michie A, Swindells MB, Orengo C, Thornton JM: **Domain assignment for protein structures using a consensus approach: characterization and analysis**. *Protein Sci* 1998, **7**(2):233-242.

17. Kawasaki H, Kretsinger RH: **Calcium-binding proteins 1: EF-hands**. *Protein Profile* 1995, **2**(4):297-490.

18. Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche BA, de Castro E, Lachaize C, Langendijk-Genevaux PS, Sigrist CJ: **The 20 years of PROSITE**. *Nucleic Acids Res* 2008, **36**(Database issue):D245-249.

19. de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N: **ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W362-365.

20. Anfinsen CB, Haber E, Sela M, White FH, Jr.: **The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain**. *Proc Natl Acad Sci U S A* 1961, **47**:1309-1314.

21. Deshpande N, Addess KJ, Bluhm WF, Merino-Ott JC, Townsend-Merino W, Zhang Q, Knezevich C, Xie L, Chen L, Feng Z *et al*: **The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema**. *Nucleic Acids Res* 2005, **33**(Database issue):D233-237.

22. Chen R, Li L, Weng Z: **ZDOCK: an initial-stage protein-docking algorithm**. *Proteins* 2003, **52**(1):80-87.

23. Gould CM, Diella F, Via A, Puntervoll P, Gemund C, Chabanis-Davidson S, Michael S, Sayadi A, Bryne JC, Chica C *et al*: **ELM: the status of the 2010 eukaryotic linear motif resource**. *Nucleic Acids Res* 2010, **38**(Database issue):D167-180.

24. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A *et al*: **ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins**. *Nucleic Acids Res* 2003, **31**(13):3625-3630.

25. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2010, **38**(Database issue):D211-222.

26. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD *et al*: **InterPro--an integrated documentation resource for protein families, domains and functional sites**. *Bioinformatics* 2000, **16**(12):1145-1150.

27. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S: **AmiGO: online access to ontology and annotation data**. *Bioinformatics* 2009, **25**(2):288-289.

28. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM *et al*: **Human protein reference database--2006 update**. *Nucleic Acids Res* 2006, **34**(Database issue):D411-414.

29. Diella F, Gould CM, Chica C, Via A, Gibson TJ: **Phospho.ELM: a database of phosphorylation sites--update 2008**. *Nucleic Acids Res* 2008, **36**(Database issue):D240-244.

30. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S: **The protein kinase complement of the human genome**. *Science* 2002, **298**(5600):1912-1934.

31. Latchman DS: **Transcription factors: an overview**. *Int J Biochem Cell Biol* 1997, **29**(12):1305-1312.

32. Karin M: **Too many transcription factors: positive and negative interactions**. *New Biol* 1990, **2**(2):126-131.

33. Blount P, Sukharev SI, Moe PC, Schroeder MJ, Guy HR, Kung C: **Membrane topology and multimeric structure of a mechanosensitive channel protein of Escherichia coli**. *EMBO J* 1996, **15**(18):4798-4805.

34. Mitchell PJ, Tjian R: **Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins**. *Science* 1989, **245**(4916):371-378.

35. Libermann TA, Zerbini LF: **Targeting transcription factors for cancer gene therapy**. *Curr Gene Ther* 2006, **6**(1):17-33.

36. Clevenger CV: **Roles and regulation of stat family transcription factors in human breast cancer**. *Am J Pathol* 2004, **165**(5):1449-1460.

37. Maestro MA, Cardalda C, Boj SF, Luco RF, Servitja JM, Ferrer J: **Distinct roles of HNF1beta, HNF1alpha, and HNF4alpha in regulating pancreas development, beta-cell function and growth**. *Endocr Dev* 2007, **12**:33-45.

38. Al-Quobaili F, Montenarh M: **Pancreatic duodenal homeobox factor-1 and diabetes mellitus type 2 (review)**. *Int J Mol Med* 2008, **21**(4):399-404.

39. Moretti P, Zoghbi HY: **MeCP2 dysfunction in Rett syndrome and related disorders**. *Curr Opin Genet Dev* 2006, **16**(3):276-281.

40. Overington JP, Al-Lazikani B, Hopkins AL: **How many drug targets are there?** *Nat Rev Drug Discov* 2006, **5**(12):993-996.

41. Warnmark A, Treuter E, Wright AP, Gustafsson JA: **Activation functions 1 and 2 of nuclear receptors: molecular strategies for transcriptional activation**. *Mol Endocrinol* 2003, **17**(10):1901-1909.

42. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Pruss M, Reuter I, Schacherer F: **TRANSFAC: an integrated system for gene expression regulation**. *Nucleic Acids Res* 2000, **28**(1):316-319.

43. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A *et al*: **Human Protein Reference Database--2009 update**. *Nucleic Acids Res* 2009, **37**(Database issue):D767-772.

44. Lienhard GE, Secemski, II, Koehler KA, Lindquist RN: **Enzymatic catalysis and the transition state theory of reaction rates: transition state analogs**. *Cold Spring Harb Symp Quant Biol* 1972, **36**:45-51.

45. Schramm VL: **Enzymatic transition-state analysis and transition-state analogs**. *Methods Enzymol* 1999, **308**:301-355.

46. Bock JR, Gough DA: **Predicting protein--protein interactions from primary structure**. *Bioinformatics* 2001, **17**(5):455-460.

47. Wu X, Zhu L, Guo J, Zhang DY, Lin K: **Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations**. *Nucleic Acids Res* 2006, **34**(7):2137-2150.

48. Sikic M, Tomic S, Vlahovicek K: **Prediction of protein-protein interaction sites in sequences and 3D structures by random forests**. *PLoS Comput Biol* 2009, **5**(1):e1000278.

49. Bahadur RP, Zacharias M: **The interface of protein-protein complexes: analysis of contacts and prediction of interactions**. *Cell Mol Life Sci* 2008, **65**(7-8):1059-1072.

50. Park SH, Reyes JA, Gilbert DR, Kim JW, Kim S: **Prediction of protein-protein interaction types using association rule based classification**. *BMC Bioinformatics* 2009, **10**:36.

51. Schelhorn SE, Lengauer T, Albrecht M: **An integrative approach for predicting interactions of protein regions**. *Bioinformatics* 2008, **24**(16):i35-41.

52. Mahdavi MA, Lin YH: **Prediction of protein-protein interactions using protein signature profiling**. *Genomics Proteomics Bioinformatics* 2007, **5**(3-4):177-186.

53. Jenuth JP: **The NCBI. Publicly available tools and resources on the Web**. *Methods Mol Biol* 2000, **132**:301-312.

54. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D: **DIP: the database of interacting proteins**. *Nucleic Acids Res* 2000, **28**(1):289-291.

55. Albrecht M, Huthmacher C, Tosatto SC, Lengauer T: **Decomposing protein networks into domain-domain interactions**. *Bioinformatics* 2005, **21 Suppl 2**:ii220-221.

56. Chang EJ, Begum R, Chait BT, Gaasterland T: **Prediction of cyclin-dependent kinase phosphorylation substrates**. *PLoS One* 2007, **2**(7):e656.

57. Guo J, Wu X, Zhang DY, Lin K: **Genome-wide inference of protein interaction sites: lessons from the yeast high-quality negative protein-protein interaction dataset**. *Nucleic Acids Res* 2008, **36**(6):2002-2011.

58. Brookes AJ: **The essence of SNPs**. *Gene* 1999, **234**(2):177-186.

59. Collins FS, Brooks LD, Chakravarti A: **A DNA polymorphism discovery resource for research on human genetic variation**. *Genome Res* 1998, **8**(12):1229-1231.

60. Shastry BS: **SNPs in disease gene mapping, medicinal drug development and evolution**. *J Hum Genet* 2007, **52**(11):871-880.

61. Evans WE, Johnson JA: **Pharmacogenomics: the inherited basis for interindividual differences in drug response**. *Annu Rev Genomics Hum Genet* 2001, **2**:9-39.

62. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM *et al*: **A second generation human haplotype map of over 3.1 million SNPs**. *Nature* 2007, **449**(7164):851-861.

63. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL *et al*: **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms**. *Nature* 2001, **409**(6822):928-933.

64. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic Acids Res* 2001, **29**(1):308-311.

65. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders**. *Nucleic Acids Res* 2005, **33**(Database issue):D514-517.

66. Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey**. *Nucleic Acids Res* 2002, **30**(17):3894-3900.

67. Wjst M: **Target SNP selection in complex disease association studies**. *BMC Bioinformatics* 2004, **5**:92.

68. Reumers J, Schymkowitz J, Ferkinghoff-Borg J, Stricher F, Serrano L, Rousseau F: **SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs**. *Nucleic Acids Res* 2005, **33**(Database issue):D527-532.

69. Kelemen A, Vasilakos AV, Liang Y: **Computational intelligence in bioinformatics: SNP/haplotype data in genetic association study for common diseases**. *IEEE Trans Inf Technol Biomed* 2009, **13**(5):841-847.

70. Clark TG, De Iorio M, Griffiths RC, Farrall M: **Finding associations in dense genetic maps: a genetic algorithm approach**. *Hum Hered* 2005, **60**(2):97-108.

71. Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH: **Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases**. *BMC Bioinformatics* 2003, **4**:28.

72. Motsinger AA, Lee SL, Mellick G, Ritchie MD: **GPNN: power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease**. *BMC Bioinformatics* 2006, **7**:39.

73. Hubley RM, Zitzler E, Roach JC: **Evolutionary algorithms for the selection of single nucleotide polymorphisms**. *BMC Bioinformatics* 2003, **4**:30.

74. **The Universal Protein Resource (UniProt) 2009**. *Nucleic Acids Res* 2009, **37**(Database issue):D169-174.

75. Guerois R, Nielsen JE, Serrano L: **Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations**. *J Mol Biol* 2002, **320**(2):369-387.

76. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L: **Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins**. *Nat Biotechnol* 2004, **22**(10):1302-1306.

77. Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data**. *Nucleic Acids Res* 2004, **32**(Database issue):D129-133.

78. Lu Z, Szafron D, Greiner R, Lu P, Wishart DS, Poulin B, Anvik J, Macdonell C, Eisner R: **Predicting subcellular localization of proteins using machine-learned classifiers**. *Bioinformatics* 2004, **20**(4):547-556.

79. Nakai K, Horton P: **PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization**. *Trends Biochem Sci* 1999, **24**(1):34-36.

80. Cordell HJ: **Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans**. *Hum Mol Genet* 2002, **11**(20):2463-2468.

81. Banzhaf W, Beslon G, Christensen S, Foster JA, Kepes F, Lefort V, Miller JF, Radman M, Ramsden JJ: **Guidelines: From artificial evolution to computational evolution: a research agenda**. *Nat Rev Genet* 2006, **7**(9):729-735.

82. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.

83. Kostich M, English J, Madison V, Gheyas F, Wang L, Qiu P, Greene J, Laz TM: **Human members of the eukaryotic protein kinase family**. *Genome Biol* 2002, **3**(9):RESEARCH0043.

84. Rochette-Egly C, Germain P: **Dynamic and combinatorial control of gene expression by nuclear retinoic acid receptors (RARs)**. *Nucl Recept Signal* 2009, **7**:e005.

85. Remenyi A, Good MC, Lim WA: **Docking interactions in protein kinase and phosphatase networks**. *Curr Opin Struct Biol* 2006, **16**(6):676-685.

86. Zhou T, Sun L, Humphreys J, Goldsmith EJ: **Docking interactions induce exposure of activation loop in the MAP kinase ERK2**. *Structure* 2006, **14**(6):1011-1019.

87. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2008, **36**(Database issue):D281-288.

88. Attwood TK, Croning MD, Flower DR, Lewis AP, Mabey JE, Scordis P, Selley JN, Wright W: **PRINTS-S: the database formerly known as PRINTS**. *Nucleic Acids Res* 2000, **28**(1):225-227.

89. Corpet F, Servant F, Gouzy J, Kahn D: **ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons**. *Nucleic Acids Res* 2000, **28**(1):267-269.

90. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD *et al*: **The InterPro database, an integrated documentation resource for protein families, domains and functional sites**. *Nucleic Acids Res* 2001, **29**(1):37-40.

91. Diella F, Cameron S, Gemund C, Linding R, Via A, Kuster B, Sicheritz-Ponten T, Blom N, Gibson TJ: **Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins**. *BMC Bioinformatics* 2004, **5**:79.

92. Evans P, Dampier W, Ungar L, Tozeren A: **Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs**. *BMC Med Genomics* 2009, **2**:27.

93. Dampier W, Evans P, Ungar L, Tozeren A: **Host sequence motifs shared by HIV predict response to antiretroviral therapy**. *BMC Med Genomics* 2009, **2**:47.

94. Bardwell AJ, Frankson E, Bardwell L: **Selectivity of docking sites in MAPK kinases**. *J Biol Chem* 2009, **284**(19):13165-13173.

95. Neduva V, Russell RB: **Linear motifs: evolutionary interaction switches**. *FEBS Lett* 2005, **579**(15):3342-3345.

96. Skrabanek L, Saini HK, Bader GD, Enright AJ: **Computational prediction of protein-protein interactions**. *Mol Biotechnol* 2008, **38**(1):1-17.

97. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H: **Predicting protein-protein interactions based only on sequences information**. *Proc Natl Acad Sci U S A* 2007, **104**(11):4337-4341.

98. Sprinzak E, Margalit H: **Correlated sequence-signatures as markers of protein-protein interaction**. *J Mol Biol* 2001, **311**(4):681-692.

99. Liu M, Chen XW, Jothi R: **Knowledge-guided inference of domain-domain interactions from incomplete protein-protein interaction networks**. *Bioinformatics* 2009, **25**(19):2492-2499.

100. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bahler J, Wood V *et al*: **The BioGRID Interaction Database: 2008 update**. *Nucleic Acids Res* 2008, **36**(Database issue):D637-640.

101. Prasad TS, Kandasamy K, Pandey A: **Human protein reference database and human proteinpedia as discovery tools for systems biology**. *Methods Mol Biol* 2009, **577**:67-79.

102. Gormley M, Dampier W, Ertel A, Karacali B, Tozeren A: **Prediction potential of candidate biomarker sets identified and validated on gene expression data from multiple datasets**. *BMC Bioinformatics* 2007, **8**:415.

103. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C *et al*: **The Gene Ontology (GO) database and informatics resource**. *Nucleic Acids Res* 2004, **32**(Database issue):D258-261.

104. Kanehisa M: **The KEGG database**. *Novartis Found Symp* 2002, **247**:91-101; discussion 101-103, 119-128, 244-152.

105. Fu W, Sanders-Beer BE, Katz KS, Maglott DR, Pruitt KD, Ptak RG: **Human immunodeficiency virus type 1, human protein interaction database at NCBI**. *Nucleic Acids Res* 2009, **37**(Database issue):D417-422.

106. Superti SV, Martins Dde S, Caierao J, Soares Fda S, Prochnow T, Zavascki AP: **Indications of carbapenem resistance evolution through heteroresistance as an intermediate stage in Acinetobacter baumannii after carbapenem administration**. *Rev Inst Med Trop Sao Paulo* 2009, **51**(2):111-113.

107. Slaughter BD, Schwartz JW, Li R: **Mapping dynamic protein interactions in MAP kinase signaling using live-cell fluorescence fluctuation spectroscopy and imaging**. *Proc Natl Acad Sci U S A* 2007, **104**(51):20320-20325.

108. Bhatnagar A, Ghauri AJ, Hope-Ross M, Lip PL: **Diabetic retinopathy in pregnancy**. *Curr Diabetes Rev* 2009, **5**(3):151-156.

109. Shi TL, Li YX, Cai YD, Chou KC: **Computational methods for protein-protein interaction and their application**. *Curr Protein Pept Sci* 2005, **6**(5):443-449.

110. Gomez SM, Choi K, Wu Y: **Prediction of protein-protein interaction networks**. *Curr Protoc Bioinformatics* 2008, **Chapter 8**:Unit 8 2.

111. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ: **The PROSITE database**. *Nucleic Acids Res* 2006, **34**(Database issue):D227-230.

112. Sonnhammer EL, Eddy SR, Durbin R: **Pfam: a comprehensive database of protein domain families based on seed alignments**. *Proteins* 1997, **28**(3):405-420.

113. Ta HX, Holm L: **Evaluation of different domain-based methods in protein interaction prediction**. *Biochem Biophys Res Commun* 2009, **390**(3):357-362.

114. McClellan J, King MC: **Genetic heterogeneity in human disease**. *Cell* 2010, **141**(2):210-217.

115. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL: **The Pfam protein families database**. *Nucleic Acids Res* 2000, **28**(1):263-266.

116. Bourne PE, Addess KJ, Bluhm WF, Chen L, Deshpande N, Feng Z, Fleri W, Green R, Merino-Ott JC, Townsend-Merino W *et al*: **The distribution and query systems of the RCSB Protein Data Bank**. *Nucleic Acids Res* 2004, **32**(Database issue):D223-225.

117. Yue P, Melamud E, Moult J: **SNPs3D: candidate gene and SNP selection for association studies**. *BMC Bioinformatics* 2006, **7**:166.

118. Krieger E, Koraimann G, Vriend G: **Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field**. *Proteins* 2002, **47**(3):393-402.

119. Huang da W, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler MW, Lane HC *et al*: **DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W169-175.

120. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Res* 1999, **27**(1):29-34.

121. Raghavachari B, Tasneem A, Przytycka TM, Jothi R: **DOMINE: a database of protein domain interactions**. *Nucleic Acids Res* 2008, **36**(Database issue):D656-661.

122. Arking DE, Chakravarti A: **Understanding cardiovascular disease through the lens of genome-wide association studies**. *Trends Genet* 2009, **25**(9):387-394.

123. Cichon S, Craddock N, Daly M, Faraone SV, Gejman PV, Kelsoe J, Lehner T, Levinson DF, Moran A, Sklar P *et al*: **Genomewide association studies: history, rationale, and prospects for psychiatric disorders**. *Am J Psychiatry* 2009, **166**(5):540-556.

124. Kronenberg F: **Genome-wide association studies in aging-related processes such as diabetes mellitus, atherosclerosis and cancer**. *Exp Gerontol* 2008, **43**(1):39-43.

125. Eglen RM, Reisine T: **The current status of drug discovery against the human kinome**. *Assay Drug Dev Technol* 2009, **7**(1):22-43.

126. Moore JH, Asselbergs FW, Williams SM: **Bioinformatics challenges for genome-wide association studies**. *Bioinformatics* 2010, **26**(4):445-455.

127. **The International HapMap Project**. *Nature* 2003, **426**(6968):789-796.

128. Maniatis N: **Linkage disequilibrium maps and disease-association mapping**. *Methods Mol Biol* 2007, **376**:109-121.

APPENDICES

Appendix A: PROSITE domains that are statistically enriched in subtypes of kinases and substrates. Columns of the table represent domain index used in Figure 2.1B, domain name, domain PROSITE ID Number as well as the kinase groups for which the domain is enriched.
http://code.google.com/p/tozerenlab/downloads/list

Appendix B: The list of domain-strings pairs used in predicting phosphorylation PPI with high specificity (SP > 0.91). DSIK: Domain string index for the kinase in PPI; DSIS: Domain string index for the substrate in PPI.
http://code.google.com/p/tozerenlab/downloads/list

VITA

Contact: Yichuan Liu, yl388@drexel.edu, 215-8820139

Education:

Drexel University, PA, USA Sep 2007~Present
Doctor of Philosophy (PhD) in Bioinformatics/Biomedical Engineering

University of Waterloo, ON, CA Jan 2004~May 2007
Bachelor of Computer Science (BCS)
Major in Computer Science, Honors; Minor in Biology

Publications:

Y. Liu, A.Tozeren. Modular composition predicts kinase/substrate interactions; requested revisions submitted to BMC Bioinformatics 2010.

Y. Liu, A.Tozeren. YiRen: A human protein-protein interaction (PPI) prediction tool based on functional domains, Bioinformatics 2010; waiting for final approval from advisor for submission

Y. Liu, A.Tozeren. Predicting the association between human SNPs and diseases based on signal strength alterations of functional domains, Biophysical Journal 2010; currently under views

Activities:

Research Assistant for the Center of Integrated Bioinformatics (CFIB)

TA for graduate course: Quantitative Systems Biology; Drexel University (Jan 2009 - May 2009)

TA for graduate online course: Short courses in Bioinformatics and Computational Biology; GPBA (June 2008 - Aug 2008)

Informatics consultant, Project: "NCBI CN, subtitle: Human Retina and Optic Nerve" under China National Program on Key Basic Research Project (973), #2004CB720300 (Aug 2006 – present)